REAL TIME BEST VIEW SELECTION IN CYBER-PHYSICAL ENVIRONMENTS

WANG YING

(B. S.), Xi'an Jiao Tong University, China

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgements

I would like to express my sincere gratitude to all those who have given me the support to complete this thesis. I want to thank the Department of Computer Science, School of Computing for giving me the opportunity to commence on this thesis and permission to do necessary research work and to use departmental facilities. I would like to especially thank my supervisor, Prof. Mohan S. Kankanhalli who has been continuously giving me a lot of guidance, encouragement, and support throughout the process of this research work.

Furthermore, I would like to thank my previous graduate research paper examiners, A/Prof. Roger Zimmermann and A/Prof. Kok-Lim Low for their valuable suggestions on improving this work. Also I would like to thank all my colleagues from Multimedia Analysis and Synthesis Lab and Dr. S. Ramanathan for their help during the time of conducting this research work.

Additionally, I want to thank my friends for photographing the experimental images and giving me suggestions on implementing the system.

Finally, I would like to give my special thanks to my parents, whose deep love and mental support enabled me to fulfill this work.

Table of Contents

Summary

With the rapid spread of the Internet, more and more people are benefitting from services such as online chatting, video conferencing, VoIP applications and distance education. Our goal is to build upon this trend and improve the Quality of Experience of remote communication systems such as video conferencing. In this thesis, we propose a novel approach towards real-time selection and acquisition of the best view of user-selected objects in remote cyber-physical environments equipped with multiple IP network cameras over the Internet. Traditional three-dimensional viewpoint selection algorithms generally rely on the availability of the 3D model of the physical environment and therefore require a complex model computation process. Therefore, they may work well in completely synthetic environments where the 3D model is available, but are not applicable for the real time communication applications in cyber-physical environments where the response time is a key issue. To address this problem, we first define a new image based metric, Viewpoint Saliency (VS), for evaluating the quality of viewpoints for a captured cyber-physical environment, and then based on this new metric, we propose a scheme for controlling multiple cameras to obtain the best view upon the user's selection. Since the Viewpoint Saliency measure is purely image-based, 3D model reconstruction is not required. And then we map the real time best view selection and acquisition problem to a "Best Quality Least Effort" task on a graph formed by available views of an object and model it as a finite cameras state transition problem for energy minimization where the quality of the view measured by VS and its associated cost serve as individual energy terms in the overall energy function. We have implemented our method and the experiments show that the proposed approach is indeed feasible and effective for real time applications in cyber-physical environments.

List of Tables

List of Figures

List of Symbols

I_v	Viewpoint entropy (Equation 2.1)
N_f	Total number of faces of the scene (Equation 2.1)
A_i	The projected area of face *i* over the sphere (Equation 2.1)
A_t	Total area of the sphere (Equation 2.1)
S	A scene (Equation 2.2)
p	A viewpoint from scene *S* (Equation 2.2)
N_{pixi}	The number of the projected pixels of face *i* (Equation 2.2)
N_{pixF}	The total number of pixels of the image (Equation 2.2)
$C(V)$	The viewpoint quality of the scene or object (Equation 2.3)
$P_i(V)$	The number of pixels corresponding to the polygon i in the image obtained from the viewpoint V (Equation 2.3)
n	The total number of polygons of the scene (Equation 2.3)
r	The total number of pixels of the image (Equation 2.3)
$U(v)$	The saliency visible from viewpoint v (Equation 2.4)
$F(v)$	The set of surface points visible from viewpoint v (Equation 2.4)
g	Mesh saliency (Equation 2.4)
v_m	The viewpoint with maximum visible saliency (Equation 2.5)
o_i	One polygon of an object or scene (Equation 2.6)

$S(o_i)$	The saliency of a polygon o_i (Equation 2.6)
N_0	The number of neighbor polygons of o_i (Equation 2.6)
JS	Jensen-Shannon divergence (Equation 2.6)
$B_{j,k}^{S}$	The j-th blob extracted from sensor s at time k (Equation 2.7)
$D(x, y)$	The gray-scale value of pixel at position *(x, y)* in the difference map (Equation 2.8)
Th_K^S	The threshold (Equation 2.8)
$H(X)$	Shannon entropy of random variable *X* (Equation 2.9)
P_i	Probability distribution (Equation 2.10)
w_i	The weight for the probability distribution P_i (Equation 2.10)
u_x	The mean of an image x (Equation 2.11, 2.12, 2.13)
o_x^2	The variance of an image x (Equation 2.11, 2.12, 2.13)
$o_{x,y}$	The covariance of image x and y (Equation 2.11, 2.12, 2.13)
I	Luminance comparison measure in SSIM (Equation 2.11)
C	Contrast comparison measure in SSIM (Equation 2.12)
S	Structure comparison measure in SSIM (Equation 2.13)
Q	Final quality score (Equation 2.17)
f	The spatial frequency of the visual stimuli (Equation 2.18)
$A(f)$	The contrast sensitivity function (Equation 2.18)
M	The mask (Equation 2.19)
I	The image (Equation 2.19)

C_m	The composite contrast map (Equation 2.19)
$C_{i,j}$	The contrast value on a perception unit (i, j) (Equation 2.20)
$p_{i,j}$	The stimulus perceived by perception unit (i, j) (Equation 2.20)
$\boldsymbol{\Theta}$	The neighborhood of perception unit (i, j) (Equation 2.20)
VS	Viewpoint Saliency (Equation 3.1, 3.2, 3.7)
p_c	The contrast descriptor (Equation 3.2, 3.3)
p_a	The projected area descriptor (Equation 3.2, 3.6)
O	A bounded object region (Equation 3.3)
N_p	The total number of perception units within the object region (Equation 3.3)
$C_{p_{ij}}$	Contrast level value of the perception unit $p_{i,j}$ obtained from the contrast map (Equation 3.3)
d	The distance measure (Equation 3.4)
W	The width of the object region (Equation 3.6)
H	The height of the object region (Equation 3.6)
M	The height of the image (Equation 3.6)
N	The width of the image (Equation 3.6)
a	The scaling factor (Equation 3.6)
w_1	The weight of contrast level descriptor p_c (Equation 3.7)
w_2	The weight of projected area descriptor p_a (Equation 3.7)
$\boldsymbol{G}$	The graph formed by available views of an object (Section 3.4.1)
V	The set of views that can be captured by all the cameras (Section 3.4.1)

E	The set of edges in graph G (Section 3.4.1)
e_i	An edge in the graph **G** (Equation 3.8)
u_i	The starting node linked by edge e_i (Equation 3.8)
t_i	Time required for moving a camera from starting node to ending node (Equation 3.8)
v_i	The ending node linked by edge e_i (Equation 3.8)
S	The set of camera states throughout best view selection (Section 3.4.1)
$\boldsymbol{E}(S_i)$	The total energy of cameras state S_i (Equation 3.9)
$\boldsymbol{E}_{quality}\ (S_i)$	The quality energy term of cameras state S_i (Equation 3.9)
$\boldsymbol{E}_{cost}\ (S_i)$	The cost energy term of cameras state S_i (Equation 3.9)
α_1	The weight of quality energy term (Equation 3.9)
α_2	The weight of cost energy term (Equation 3.9)
A_j	Current search area for a camera (Algorithm 3.1)
$A_j^{'}$,	New search area for a camera (Algorithm 3.1)

1. Introduction

In recent years, more and more emphasis has been laid on improving the QoE (Quality of Experience) when designing new multimedia systems or applications. Quality of experience, also sometimes known as "Quality of User Experience", is a multi-dimensional construct of perception and behavior of a user, which captures his/her emotional, cognitive and behavioral responses, both subjective and objective while using a system [73]. It indicates the degree of a user's satisfaction. It is related to but is different from the Quality of Service (QoS) concept, which refers to an objective system performance metric, such as the bandwidth, delay, and packet loss rate of a communication network [11].

Cyber-physical systems are systems featuring a tight combination of, and coordination between, the system's computational, sensing, communication, control and physical elements. Ideally, these functions provided by cyber-physical systems that support human activities in everyday life should allow them to interact with humans adaptively according to context, such as the situation in the real world and each human's individual characteristics. With the advances in communication, control and sensing technologies, various information through different types of media, i.e. video, audio and image, can be presented to users in real time, not only making it possible for cyber-physical systems to support intellectual activities such as conferencing, surveillance and interactive TV, but also opening great possibilities of achieving intelligent functions to improve their QoE. In cyber-physical environments, where rapid user interactions are enabled, one useful intelligent function would be providing the user with the best view of his/her own object(s) of interest, whereas the meaning of "the best" could vary from object to object and from one person to another. For example, in a multimodal conferencing application, a user may want to better see the desk at the remote environment.

Especially, in the application of video conferencing, surveillance or interactive TV, where the systems usually contain multiple sensors such as video cameras to capture different views of monitored scene, it is useful to decide the best viewpoint of objects included in the monitored scene. Therefore, it is essential to develop a fast viewpoint quality assessment algorithm which can accomplish the task in real time.

However, there is only limited work that has been done in this area. Previous best view(s) selection algorithms [3, 63, 64, 65, 66] either require prior knowledge of the geometry of the scene and objects and relies on the availability of the 3D model of them [3, 63, 64, 66] or assume a fixed view such as the side view as the best view of an object [65]. Selections are usually made assuming that all the possible views can be captured by cameras. This is useful in a completely synthetic computer graphics environment but it is not applicable to cyber-physical systems such as surveillance or video conferencing systems which include fixed number of sensors and require real-time processing.

The study of visual attention is related to a few fields, including biology, psychology, neuro-psychology, cognitive science and computer vision. The research on attention began with William James, who first outlined a theory of human attention [23]. After him, more and more researchers joined in this area. So far, although the attention mechanism of human being has not been completely understood, some proven conclusion can be used to guide its application.

Previous research in 2D feature of image has shown that the contrast information can provide a fast and effective methodology to semantic image understanding [49]. Contrast-based visual attention analysis aims to explore semantic meanings of image region through a simple low level feature – contrast [49]. Other features, such as color, texture, and shape were adopted to build human visual attention models such as Itti visual attention model [22], however, were proved by Ma et al. [49] to be not as effective as contrast. Meanwhile, contrast, as a key factor in assessing vision, is often used in clinical settings for visual acuity measurement [49], and in reality, objects

and their surroundings are of varying contrast. Therefore, the relationship between visual acuity and contrast allows a more detailed understanding of human visual perception [28]. Hence we contemplate that some simple 2D features of an image such as contrast information (see section 2.7, section 3.3.1) may be able to provide us with an opportunity to evaluate viewpoint quality in 2D space.

In this work, our goal is to improve the QoE of real time steaming applications for video conferencing and distance communication in cyber physical environments by making use of multimedia sensing and computing. We aim to improve the users' experience by allowing them to select objects of interest in a remote cyber physical environment equipped with multiple cameras. Figure 1.1 illustrates this idea of our work.

As it is shown in Figure 1.1, the best view selection problem in this work is stated as follows: assume that a user is connected to a remote cyber-physical environment which has several video cameras. The user would like to obtain a good view of some object(s) of interest in the remote environment. The proposed algorithm will help the viewers to automatically obtain the best view of the object(s) in real time. The object(s) covered here include general objects, human being and the algorithm is able to detect the slow motion of objects of interest and make adaptive responses.

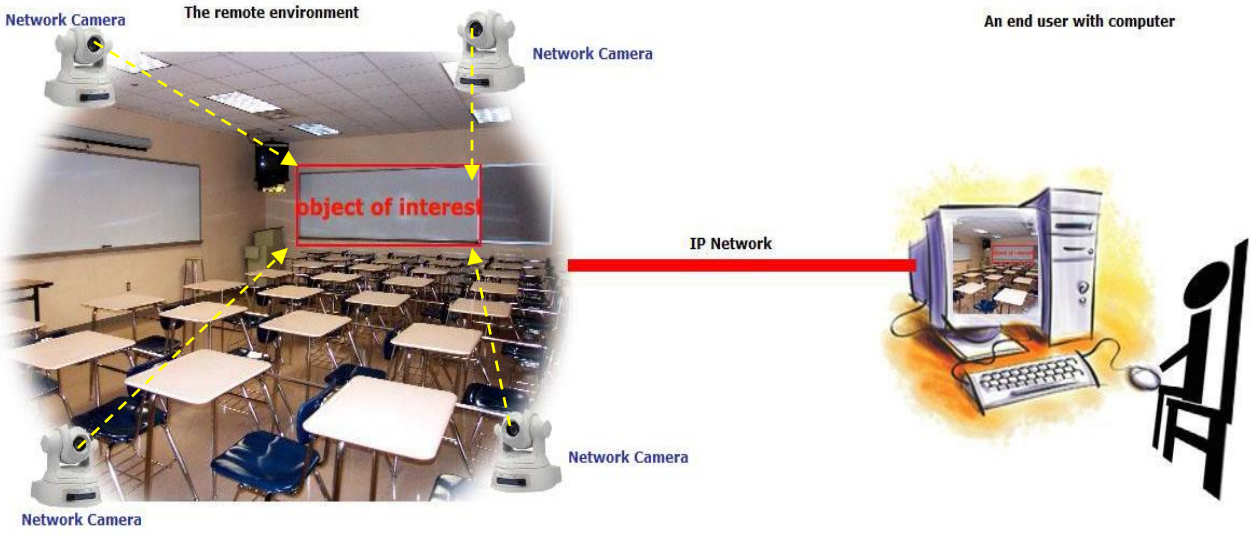


Figure 1.1 Illustration of the best view selection problem

To make best view acquisition feasible for real time streaming applications such as video conferencing in cyber-physical environments, we first propose a novel image-based metric, *Viewpoint Saliency* (VS), for evaluating the quality of different viewpoints for a given object. This measure is fast and can eliminate 3D model reconstruction. Using VS, best views of user selected objects can be acquired through feedback based camera control and delivered via Internet in real time. The new image based "best viewpoint" measure has been tested with general objects and humans. We also pose the real time best view computation problem as a "Best Quality Least Effort" task performed on a graph formed by available views of an object, and then formulate it as a unified energy minimization problem where the quality of the view measured by VS and its associated cost incurred by cameras' movements are represented by two energy terms. Finally, to demonstrate our algorithm, we provide various experiment results with our implemented VC++ based system.

The contributions of this thesis are as follows: first, an image based viewpoint evaluation metric, *Viewpoint Saliency*, is developed and tested; second, an energy minimization based camera control algorithm is proposed for acquiring the best view(s) of object(s) of interest to with the goal of "Best Quality Least Effort"; third, a system which supports remote best view selection and acquisition via Internet is implemented and tested with four IP network cameras on VC++ platform.

The rest of this thesis is organized as follows: chapter 2 is the detailed review of previous related work. Chapter 3 gives the details of the proposed approach. Chapter 4 presents the system demonstration as well as the analysis of results. Chapter 5 concludes the thesis with a summary of the overall work and major contributions as well as a brief outline of future work.

2. Related Work

The research of real time best view selection in cyber physical environment is related to eight major research areas in multimedia research, namely, Internet supported tele-operation and communication, three dimensional viewpoint selection and evaluation, multi-camera system, information theory, visual attention analysis, visual quality assessment, the contrast feature of images, template matching and segmentation. The literature survey of this work was done with a focus on the above eight domains, and the following is a detailed review of previously most relevant work.

2.1. Internet supported tele-operation and communication

In the field of internet robotics, Mosher [46] at GE demonstrated a complex two arm tele-operator with video camera in the 1960s. The Mercury Project developed by Goldberg et al [18] was the first system to permit Internet users to remotely view and manipulate a camera through robots over the WWW. The control of networked robotic cameras [59, 60] were also studied for remote observation applications such as nature observation, surveillance and distance learning

In the area of video conferencing via Internet, Liu et al [33] combined a fixed panoramic camera with robotic pan-tilt-zoom camera for collaborative video conferencing based on WWW. They address the frame selection problem by partitioning the solution space into small non-overlapping regions. They estimate the probability that each small region will be viewed based on the frequency that this region intersects with user requests. Based on the probability distribution, they choose the optimum frame by minimizing the discrepancy in the probability based estimation.

Although most of previous work in internet supported tele-operation and communication addressed the problem of frame selection for collaboratively controlled robotic camera, none of

them have looked into the content of one specific camera view for "best view" selection. Knowing that Internet and WWW can provide a good platform for the "best view" selection system to run, we still need to develop feasible viewpoint quality evaluation and cameras control algorithm for system implementation.

2.2. Three-dimensional viewpoint selection and evaluation

2.2.1. Viewpoint entropy

Vazquez et al [66, 67] was inspired by the theory of Shannon's information entropy and defined viewpoint entropy as the relative area of the projected faces of an object over the sphere of directions centered at viewpoint v. The mathematical definition of viewpoint entropy was given as

$$I_v = -\sum_{i=0}^{N_f} \frac{A_i}{A_t} log \frac{A_i}{A_t} \qquad (2.1)$$

where N_f is the total number of faces of the scene, A_i is the projected area of face i over the sphere, A_0 represents the projected area of background in the open scene, and A_t is the total area of the sphere. The maximum viewpoint entropy is obtained when a certain viewpoint can see all the faces with the same projected area. The best viewpoint is defined as the one that has the maximum entropy.

Based on viewpoint entropy, a modified measure----orthogonal frustum entropy [68] was introduced for obtaining good views of molecules. It is a 2D based version of previous viewpoint entropy measure. The orthogonal frustum entropy of a point *p* from a scene *S* is defined as:

$$I_O(S,p) = -\sum_{i=0}^{N_f} \frac{N_{pixi}}{N_{pixF}} * log \frac{N_{pixi}}{N_{pixF}} \qquad (2.2)$$

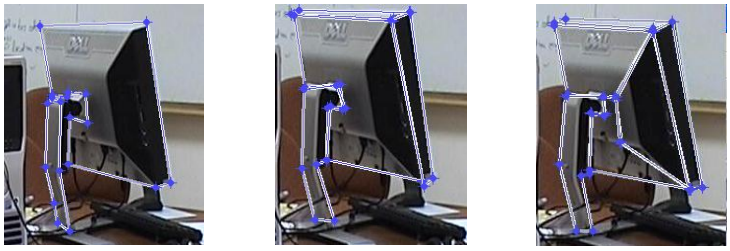
where N_{pixi} is the number of the projected pixels of face i, and N_{pixF} is the total number of pixels of the image. This measure is appearance-based in the sense that it only measures what we can really see. This means that we will apply it to the objects that project at least one pixel on the screen, which are perceivable by an observer. Good views of molecules were defined by the following criterion:

(1) views with high orthogonal entropy of single molecules.

(2) views with low orthogonal entropy of arrangements of the same molecule.

Centred around viewpoint entropy theory, there are a number of algorithms that were developed [66, 67, 68, 69, 70] for various applications, including image based rendering [67] and automatic indoor scene exploration [69]; and for improving the performance of the algorithms [69, 70].

Though viewpoint entropy is proved to be an effective measure of the viewpoint quality in a completely synthetic environment, which is useful for computer graphics based research, it is almost impossible to adopt it for real time applications because of its limitations in algorithm robustness and computation cost. The main drawback of viewpoint entropy is that it relies on the 3D model of an object; additionally, it depends on the polygonal discretisation of object's faces [16, 63]. A heavily discretised region will boost the value of viewpoint entropy, and hence the measure favors small polygons more than large ones.



a **b** **c**

Figure 2.1 Different segmentation of faces of the computer monitor. a. Two faces. b. Four faces, c. Six faces.

Table 2.1 Viewpoint entropy of the same image when different numbers of faces are segmented. (I_a, I_b, I_c are corresponding to image a, b, c in Figure 2.1)

	2 Faces	4 Faces	6 Faces
View Entropy	I_a = 0.0709	I_b = 0.0842	I_c = 0.0956

Figure 2.1 and Table 2.1 demonstrate the behavior of viewpoint frustum entropy [68] under different granularities of segmenting object's faces. It can be seen that viewpoint entropy heavily depends on the segmentation of faces of the object in the image.

Table 2.1 is the viewpoint entropy computed of the same image under different segmentation schemes, it is shown that viewpoint entropy is largely related to the number of the faces segmented for an object in the image.

2.2.2. Heuristic measure

Barral et al [3, 62] introduced a method for visual understanding of a scene by efficient automatic movement of a camera. The purpose of this method is to choose a trajectory for a virtual camera based on the 3D model of the scene, allowing the user to have a good knowledge of the scene. The method is based on a heuristic measure for computing the quality of a viewpoint of a scene. It is defined as follows:

$$C(V) = \frac{\sum_{i=1}^{n} \left\lceil \frac{P_i(V)}{P_i(V)+1} \right\rceil}{n} + \frac{\sum_{i=1}^{n} P_i(V)}{r} \qquad (2.3)$$

where V is the viewpoint, C(V) is the viewpoint quality of the scene or object, $P_i(V)$ is the number of pixels corresponding to the polygon i in the image obtained from the viewpoint V, r is the total number of pixels of the image (resolution of the image), n is the total number of surfaces in the scene. In this formula, $\lceil x \rceil$ denotes the smallest integer, greater than or equal to x. It is

observed that the first term in (2.3) gives the fraction of visible surfaces with respect to the total number of surfaces, while the second term is the ratio between the projected area of the scene (or object) and the screen area (thus, its value is 1 for closed scene). The heuristic considers a viewpoint to be good if it minimizes maximum angle deviation between direction of view and normals to the faces and give s a high amount of details.

2.2.3. Mesh saliency

Lee et al. [40] introduced the measure of mesh saliency for achieving salient viewpoint selection. They borrowed the idea of Itti et al. [22] (refer to section 2.5. visual attention analysis) of computing saliency for 2D images and developed their own method to compute saliency of 3D meshes. Mesh saliency is formulated in terms of the mean curvature used with the center-surround mechanism. Based on the Mesh saliency, they developed a method for automatically selecting viewpoint so as to visualize the most salient object features. Their method selects the viewpoint that maximizes the sum of saliency for visible regions of the object.

For a given viewpoint *v*, let *F(v)* be the set of surface points visible from v , and let *g* be the mesh saliency. The saliency visible from *v*, denoted as *U(v)*, is computed as:

$$U(v) = \sum_{x \in F(v)} g(x) \qquad (2.4)$$

Then the best view, i.e., the viewpoint with maximum visible saliency v_m is defined as:

$$v_m = \underset{v}{argmax}\ U(v) \qquad (2.5)$$

Based on above definition, a gradient-descent-based optimization heuristic was adopted to help selecting good viewpoints.

2.2.4. Viewpoint information channel

Feixas et al. [16] introduced an information channel $V \rightarrow O$ between the random variables V and O, which respectively represent a set of viewpoints and the set of polygons of an object. They defined a "goodness" measure of a viewpoint and a similarity measure between two views, both are based on the mutual information of this channel, where the similarity between two views are measured by Jensen-Shannon divergence (JS-divergence). Based on this definition, they presented a viewpoint selection algorithm to find the minimal representative set of m views for a given object or scene by maximizing their JS-divergence (see section 2.4, formula (2.10)).

They also introduced a measure of mesh saliency by evaluating the average variation of JS-divergence between two polygons of an object. The saliency of a polygon is defined as

$$S(o_i) = \frac{1}{N_0}\sum_{j=1}^{N_0}(JS(p(V|o_i), p(V|o_j)) \geq 0 \qquad (2.6)$$

where o_j is a neighbor polygon of o_i, N_0 is the number of neighbor polygons of o_i, and the conditional probabilities are respectively weighted by $\frac{p(o_i)}{p(o_i)+p(o_j)}$ and $\frac{p(o_j)}{p(o_i)+p(o_j)}$.

2.2.5. Other related work

Apart from above past work on the definitions of best viewpoint in 3D environment, there are still a number of works that are related to viewpoint selection in three-dimensional space. The following is a brief summary of selected ones.

Moreira et al. [48] developed a model for estimating the quality of multi-views for visualization of urban rescue simulation. Their quality measure is a function of visibility, relevance, redundancy and eccentricity of the entities represented in the set of selected views. The problem

is formalized as an optimization problem to find the optimal multiple viewpoints set with appropriate view parameters that describes the rescue scenario with better quality.

Deinzer et al. [12] deals with an aspect of active object recognition for improving the classification and localization results by choosing optimal next views at an object. The knowledge of "good" next views at an object is learned automatically and unsupervised from the results of used classifier based on the eigen space approach. Methods of reinforcement learning were used in combination with numerical optimization. Though their results show that the approach is well suited for choosing optimal views at objects, however, the experiments were merely based on synthetically generated images.

Vaswani and Chellappa [65] introduced a system for selecting a single best view image chip from an IR video sequence and compression of the chip for transmission. In their work, an eigen space is constructed offline using different views (back, side and front) of the army tanks, and an assumption was made that the side view is the best view since it has most of the identifying features.

Massios and Fisher [44] proposed to evaluate the desirability of viewpoints using the weighted sum of the visibility and quality criteria. The visibility criterion maximizes the amount of occlusion plane voxels that are visible from the new viewpoint. The quality criterion maximizes the amount of low quality voxels that are visible from the new viewpoint. Both of these criteria were defined as a function of viewing direction.

There are also a few relevant works on determining the next best view [10, 35]. Low et al. [35] present an efficient next-best-view algorithm for 3D reconstruction of indoor scenes using active range sensing. To evaluate each view, they formulate a general view metric that can include many real-world acquisition constraints (i.e., scanner positioning constraints, sensing constraints, registration constraints) and quality requirements (i.e., Completeness and surface sampling quality, on the resulting 3D model).

Although previous measures work nicely in synthetic computer graphics environments where the 3D model of the object or the scene is available, either the computational complexity incurred by 3D model reconstruction or the required geometrical discretisation of the scene makes these approaches almost impossible to achieve in real time. And therefore, none of them are applicable in cyber-physical environments.

2.3. Multi-camera system

Multi-camera system, though having challenges such as view registration and object recognition, has the advantage of revealing more details of the monitored scene. In this section, previous work in multi-camera system is reviewed.

Zabulis et al. [78] presented an algorithm for constructing the environment from images recorded by multiple calibrated cameras. They propose an operator that yields a measure of the confidence of the occupancy of a voxel in 3D space given strongly calibrated image pair (I_1, I_2). The input of this measure is a world point $p \in R^3$, and the outputs are a confidence score *s(p)* (strength) and a 3D unit normal *k(p)* (orientation). Increasing the number of cameras can improve the accuracy of stereo, because it enhances the geometrical constraints on the topology of the corresponding pixels. In order to deal with multiple cameras, they extend the operator for a tuple of cameras, where M binocular pairs are defined.

Snidaro et al. [62] introduced an outdoor multi-camera video surveillance system operating under changing weather conditions. A new confidence measure, Appearance Ratio (AR) is defined to automatically evaluate the sensor's performance for each time instant. By comparing their ARs, the system can select the most appropriate cameras to perform specific tasks. When redundant

measurements are available for a target, the AR measures are used to perform a weighted fusion of them. The definition of AR is given as follows:

Given the frame I_k extracted from sensor s at time k, the threshold Th_K^S used to binarize the difference map D obtained as the absolute difference between the current frame F and a reference image, and let $B_{j,k}^S$ be the j-th blob extracted from sensor s at time k, then the Appearance for that blob is defined as

$$Appearance\left(B_{j,k}^S\right) = \frac{\sum_{x,y \in B_{j,k}^S} D(x,y)}{|B_{j,k}^S|} \qquad (2.7)$$

where $|B_{j,k}^S|$ is the number of pixels of the blob $B_{j,k}^S$. Normalizing with respect to the threshold, the Appearance Ratio (AR) is obtained:

$$AR\left(B_{j,k}^S\right) = \frac{Appearance\ \left(B_{j,k}^S\right)}{Th_K^S} \qquad (2.8)$$

In (7), *D(x,y)* is the gray scale value of the pixel at position *(x, y)* in the difference map. The reference image mentioned in the definition can be the previous frame or an updated background image.

The appearance of a blob is the average value of the blob's pixels in the difference map. The AR is a normalization to allow cross comparisons between sensors. The higher a blob's AR value for a given sensor, the more visible is the corresponding target for that sensor, and the more likely that the segmentation has been correctly performed yielding accurate measures (dimensions, area, centroid coordinates of the blob, etc.).

To overcome the difficulties such as view registration in multi-camera systems, Li et al. [36] present an approach to automatically register a large set of color images to a 3D geometric model. This approach constructs a sparse 3D model from the color images using a multi-view geometry reconstruction. In this approach, they first project special light onto the scene surfaces to increase

the robustness of the multi-view geometry reconstruction, and then the sparse model is approximately aligned with the detailed model. The registration is refined by planes found in the detailed model and finally, the registered color images are mapped to the detailed model using weighted blending. The major contribution of this work is the idea of establishing correspondence which is essential in view registration among color images instead of directly finding correspondences between 2D and 3D spaces.

Multiple camera systems challenge traditional stereo algorithms in many issues including view registration, selection of commonly visible image parts for matching, and the fact that surfaces are imaged differently from different viewpoints and poses. On the other hand, multiple cameras have the advantage of revealing occluded surfaces and covering larger areas. Therefore approaches that can overcome the challenges in multi-camera systems and fully utilize its advantage will make real time best view selection feasible in cyber-physical environments.

2.4. Information theory

Previously, there were a number of best viewpoint definitions in three dimensional spaces (see section 2.2) that were developed based on information theory. In this chapter, we first review the theoretical foundation of information theory and then we summarize some information theory based approaches.

Several definitions of best view such as viewpoint entropy [66], Kullback-Lebler Distance [64], have adopted information theory as their theoretical foundation. In information theory, the Shannon entropy [5] of a discrete random variable X with values in the set $\{a_1, a_2, \dots, a_n\}$ is defined as

$$H(X) = -\sum_{i=1}^{n} p_i log p_i \qquad (2.9)$$

where $p_i = \Pr[X = a_i]$, the logarithms are taken in base 2 and $0log0 = 0$ for continuity. As $-logp_i$ represents the information associated with the result a_i, the entropy gives the average information or the uncertainty of a random variable.

Additionally, Shannon's information theory is used for visual saliency computation based on "information maximization": (1) a model of bottom-up overt attention [7] is proposed based on the principle of maximizing formation sampled from a scene; (2) a proposal for visual saliency computation within the visual cortex [8] is put forth based on the premise that localized saliency computation serves to maximize information sampled from one's environment. A detailed explanation of visual saliency will be given in section 2.5.

The definitions of viewpoint information channel [16] and mesh saliency [16] by Feixas M.et al. were based on Jensen-Shannon divergence. In probability theory and statistics, the Jensen-Shannon divergence (JS-divergence) [5] is a popular method of measuring the similarity between two probability distributions. A more general definition, allowing for the comparison of more than two distributions, is given by

$$JS(P_1, P_2, \dots, P_i) = H(\sum_{i=1}^{n} w_i P_i) - \sum_{i=1}^{n} w_i H(P_i) \geq 0 \qquad (2.10)$$

where $w_1, w_2, \dots, w_n$ are the weights for the probability distributions $P_1, P_2, \dots, P_n$ and $H(P)$ is the Shannon entropy for distribution P. And for two distribution case, $w_1 = w_2 = \frac{1}{2}$.

Previously, information theory is adopted in developing the quality measures of a viewpoint and computing saliency in visual attention analysis. Reviewing the information theory and its relation to these approaches has provided guidance in developing the new image based viewpoint quality evaluation measure in this work.

2.5. Visual attention analysis

The analysis of visual attention, which are related to a few fields, including biology, psychology, neuro-psychology, cognitive science and computer vision, is essential for understanding the relationship between human's perception and cognition. Although the attention mechanism is not completely understood yet, some proven conclusions can be used to guide its applications. In this section, various computational visual attention models as well as selected relevant studies on visual attention analysis are reviewed.

2.5.1. Visual attention models

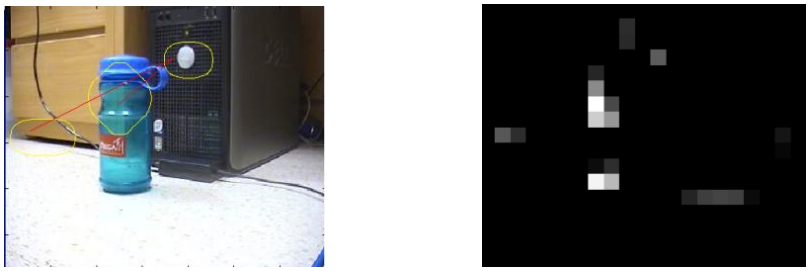
There are a number of works that have been done in this domain. Previously, many computational visual attention models have been proposed for various applications [1, 22, 32, 45, 49, 50, 55, 65, 76]. Amongst them, well known ones such as Ahmad's model [1], Niebur's model [50] and itti's model [22] are reviewed here.

A well known computational visual attention model VISIT [1]was proposed by Ahmad in 1991, which is considered to be more biologically plausible [49] than Itti's model [22]. VISIT consists of a gating network which corresponds to the pulvinar (*Medical word. The posterior medial part of the posterior end of the thalamus. It is involved in visual attention, suppression of irrelevant stimuli and utilizing information to initiate eye movements [80]*.) and its output, the gated feature maps, corresponds to the areas V4, IT and MT of the optic nerve; a priority network corresponding to the superior colliculus; frontal eye field and posterior parietal areas; a control network corresponding to the posterior parietal areas, and a working memory corresponding to the prefrontal cortex.

Niebur [50] indicated that the so-called "focus of attention" scans the scene both in the form of a rapid, bottom-up, saliency-driven and task-independent manner and in a slower, top-down, volition-controlled and task-dependent manner.

Itti et al. [22] proposed a saliency based visual attention model for rapid scene analysis. Itti's model was based on a saliency map, which topographically encodes conspicuity (or saliency) at every location in the visual input. In primates, such a saliency map is believed to be located n the posterior parietal cortex as well as in the various visual maps in the pulvinar nuclei of the thalamus. An example of Itti's saliency map is shown in Figure 2.2 below.

Their model is biologically-inspired, and is able to extract local features such as color, intensity and orientation of the input image, and construct a set of multi-scale neural "feature maps". All feature maps are then combined into a unique scalar "salience map" which encodes the saliency of a location in the scene irrespectively of the particular feature which made this location as conspicuous. In the end a Winner-Take-All competition is employed to select the most conspicuous image locations as attended points.



a **b**

Figure 2.2 Salient locations and saliency map
a. Salient locations of an image, b. Saliency map

As it is shown above, the saliency map is a function $f(x, y) \rightarrow [0,1]$, i.e., it maps every pixel to a value between 0 and 1, indicating its conspicuity in human being's perception. In comparison with Itti's saliency map, other notations of saliency map [25, 32, 49, 76] have been proposed for visual attention analysis for different purposes.

2.5.2. Visual attention based research

Visual attention is proved to be efficient in various domains of research including, image and video analysis and processing, computer graphics and computer vision.

Many have proposed to incorporate visual attention factor in objective image quality assessment [25, 39, 75] in the sense that noise will appear to be more disturbing to humans in the salient regions. Works related to image quality assessment will be reviewed in the section 2.6. For video quality assessment, Oprea et al. [53] proposed an embedded reference-free video quality metric based on salient region detection. The salient regions are estimated using the key elements that attract attention: color contrast, object size, orientation and eccentricity.

In computer graphics and vision domain, Mata et al. [47] proposed an automatic technique that makes use of the information obtained by means of a visual attention model for guiding the extraction of a simplified 3D model. Lee et al. [32] presented a real-time framework for computationally tracking objects visually attended by the users while they are navigating the interactive virtual environments. This framework can be used for perceptually based rendering without employing an expensive eye tracker, such as providing the depth-of filed effects and managing the level of detail in virtual environments.

Additionally, Li et al. [38] demonstrated an application which provides contextual advertising platform for online image service, called ImageSense, which is based on visual attention detection. Unlike most current ad-networks which treat image advertising as general text advertising by displaying relevant ads based on the contents of the Web page, ImageSense aims to embed advertisements with suitable images according to its contextual relevance to the Web page at the position where it is less intrusive and disturbing.

Knowing that the “Best Views” of object(s) has strong relationship with human visual system and human perception, reviewing previous work in visual attention analysis has helped us to understand human visual system and various approaches to content based analysis of images and their relationships to human perception.

2.6. Visual quality assessment

Best view selection using 2D features based viewpoint quality evaluation requires finding out the relationship between viewpoint quality and 2D information of images. Hence, it is important to learn about commonly used quality assessment methods. In the following paragraphs, selected works on image quality assessment are reviewed.

2.6.1. Subjective method

Radun et al. [56] used an interpretation-based quality (IBQ) estimation approach, which combines qualitative and quantitative methodology, to obtain a holistic description of subjective image quality. Their result of the test shows that the subjective effect of sharpness varies with different image content, suggesting sharpness manipulations might have different subjective meanings in different image content, which can be conceptualized as the relation between detection and preference.

The IBQ method enables simultaneous examination of psychometric results and detection, subjective preferences. IBQ method consists of qualitative part and psychometric image-quality measurement part. In their study, the qualitative part was the free sorting of the pictures, where

observers sorted each of the contents according to the similarity perceived in these pictures. They then described and evaluated the groups they had formed. The observers were not told how they should evaluate the pictures, just that they were all different. The psychometric method used was magnitude estimation of the variable sharpness to find out how the observers detected the changes in the pictures.

Their study shows that IBQ estimation is suitable and useful for image-quality studies, since a hybrid qualitative and quantitative approach can offer relevant explanations for differences seen in magnitude estimations. It helps to understand the subjective quality variations occurring in the different image contents. This is important for interpreting the results of the subjective image-quality measurements, especially in the case of high image quality, where the differences between image quality levels are small.

2.6.2. Objective method

There are various objective image quality metrics, but the most widely used image quality metrics are the mean square error (MSE) and the derived peak signal to nose ratio (PSNR) [25].These methods are simple but rather inconsistent with the subjective image quality assessments.

Other simple but far more accurate metric is structural similarity (SSIM) index [75]. SSIM metric compares local patterns of pixel intensities and therefore takes Human Visual System (HVS) into account and is highly adapted for gathering structural information. The definition of SSIM is as follows:

Let x and y be two image patches extracted from the same position in the compared images.

Let (u_x, u_y), (o_x^2, o_y^2) and $o_{x,y}$ be the mean, variance and covariance of x and y, then the luminance *I(x, y)*, and contrast *C(x, y)* and Structure *S(x, y)* comparison measures are as follows:

$$I(x,y) = \frac{2u_x u_y + C_1}{u_x^2 + u_y^2 + C_1} \quad (2.11)$$

$$C(x,y) = \frac{2o_x o_y + C_2}{o_x^2 + o_y^2 + C_2} \quad (2.12)$$

$$S(x,y) = \frac{o_{x,y} + C_3}{o_x o_y + C_3} \quad (2.13)$$

where $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, and $C_3 = C_2/2$ are small constants, L is the pixel value dynamic range, and K_1, K_2 << 1 are constants. If we consider SSIM index can be calculated as the product of above given measure, then it is calculated as follows:

$$SSIM(x,y) = \left(\frac{2u_x u_y + C_1}{u_x^2 + u_y^2 + C_1}\right)\left(\frac{2o_{x,y} + C_2}{o_x^2 + o_y^2 + C_2}\right) \quad (2.14)$$

Usually the mean SSIM index (MSSIM) is used to evaluate the overall image quality:

$$MSSIM(X,Y) = \frac{1}{N_x N_y}\sum_{x,y=1}^{N_x,N_y} SSIM(x,y) \quad (2.15)$$

where *X* and *Y* are images being compared (reference and distorted), and N_x, N_y are the picture dimensions.

SSIM metric evaluates visual quality based on the premise that the human visual system (HVS) has evolved to process structural information from natural images; hence, a high-quality image is the one whose structure closely matches the original. To this end, SSIM employs a modified measure of spatial correlation between the pixels of the reference and test images to quantify the degradation of an image's structure. Despite its simple mathematical form, SSIM objectively predicts subjective ratings as well as more sophisticated quality assessment algorithms [59]. Furthermore, SSIM's simplicity has intrigued researchers investigating how the HVS evaluates quality [75].

SSIM evaluates perceptual quality using three spatially local evaluations: mean, variance, and cross-correlation. Rouse et al. [57] investigated how the three SSIM components contribute to its quality evaluation of common image artefacts. A gradient analysis was used to illustrate the value of SSIM cross correlation component over the other two components.

The visual attention is not taken into account in SSIM for image quality assessment. Fliegel [25] presented an approach to predict perceived quality of compressed images incorporating real visual attention coordinates by implementing gaze information into image quality assessment system. The idea lies in that the artifacts are more disturbing to a human observer in the region with higher saliency than in other parts of an image. The smoothed visual attention map, which is calculated for each test image and each observer, is used to incorporate the visual attention into MSSIM index to get the visual attention weighted SSIM (ASSIM):

$$ASSIM(X,Y) = \frac{1}{N_x N_y}\sum_{x,y=1}^{N_x,N_y} AM(x,y)SSIM(x,y) \qquad (2.16)$$

where *AM* is the smoothed average visual attention map, which is obtained by directly tracking the eye movements of observers , and N_x, N_y are the picture dimensions.

Wei et al. [76] gave another human visual system based model for objective image quality estimation. They claim luminance is the first stimulus to the HVS. Then, the complexity of changes and details can be described as frequency information, which is the second stimulus; And the third most important information to visual image quality are edges which serve as the third stimulus. Hence, their final quality scores *Q* of images is implemented as follows:

$$Q = a + a_1Q_1 + a_2Q_2 + a_3Q_3 \qquad (2.17)$$

where Q_1, Q_2, Q_3 are information of luminance, Contrast Sensitivity Function (CSF) (equation (2.18), see section 2.7 for details) in the frequency domain and edge information respectively. And a_1, a_2, a_3 are constant values.

Natural images convey useful information to humans. Rose et al. [58] further investigated image utility assessment and its relationship with image quality assessment. They claim that current quality assessment algorithms implicitly assess utility insofar as an image that exhibits strong perceptual resemblance to a reference is also of high utility. However, a perceived quality score cannot predict a perceived utility score: a decrease in perceived quality may not affect the perceived utility [58]. They proposed an algorithm, referred to as the natural image contour evaluation (NICE), for assessing image utility. NICE conducts a comparison of the contours of a test image to those of a reference image across multiple image scales to score the test image. It is capable of predicting perceived utility scores and has demonstrated a viable departure from traditional quality assessment algorithms that incorporate energy-based approaches. A new metric [79] was recently proposed to more fully exploit the attributes of visual attention information.

Although image quality metrics link human visual attention with the assessment of image quality attributes such as sharpness or brightness, none of them address the problem of assessing the quality of viewpoints captured in images. Therefore, it is challenging yet interesting to develop a purely image based viewpoint quality evaluation metric to facilitate real time best view selection in cyber-physical environments.

2.7. The contrast feature

As the Contrast feature of images are of great importance to image quality assessment, it is interesting to find out how contrast feature can be used in viewpoint quality assessment, further more in real time best view selection. In this chapter, first, the basic knowledge of contrast information is introduced; and then, selected works based on contrast information are reviewed.

2.7.1. Basics of contrast information

In visual perception of the real world, contrast is determined by the difference in the color and brightness of the object and other objects within the same field of view. Because the human visual system is more sensitive to contrast than absolute luminance, we can perceive the world similarly regardless of the huge changes in illumination over the day or from place to place [28].

Contrast sensitivity is sometimes called visual acuity [28]. Mannos and Sakrison [42] proposed a model of the human contrast sensitivity function (CSF). The contrast sensitivity function tells us how sensitive we are to the various frequencies of visual stimuli. If the frequency of visual stimuli is too high we will not be able to recognize the stimuli pattern any more. Imagine an image consisting of vertical black and white stripes: if the stripes are very thin (i.e. a few thousand per millimeter), we will be unable to see individual stripes. All that we will see is a gray image. If the stripes then become wider and wider, there is a threshold width, from which on we are able to distinguish the stripes. The contrast sensitivity function proposed by Manos and Sakrison is

$$A(f) = 2.6 \cdot (0.0192 + 0.114 \cdot f) \cdot e^{-(0.114 \cdot f)^{1.1}} \qquad (2.18)$$

f in equation (2.18) is the spatial frequency of the visual stimuli given in cycles/degree. The function has a peak of value 1 approximately at f = 8.0 cycles/degree, and is meaningless for frequencies above 60 cycles/degree. The following figure (Figure 2.3) shows the contrast sensitivity function $A(f)$.

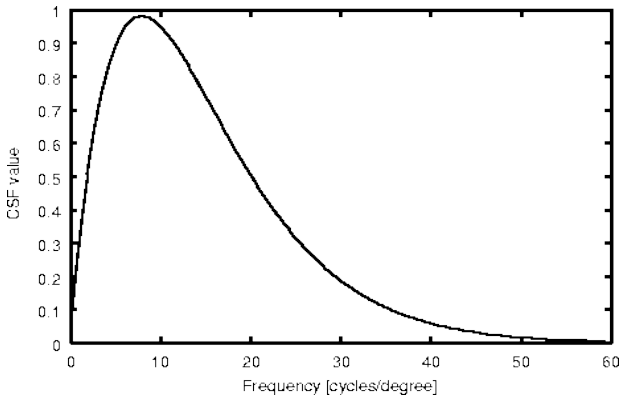


Figure 2.3 Contrast sensitivity function

2.7.2. Image contrast feature based research

Eli Peli [15] investigated various definitions of contrast in complex images. Khwaja A. A. et al. [26] presented a novel approach to manipulate an image in its contrast domain. An iterative algorithm is introduced for the reconstruction of natural images merely based on their contrast information. The solution is neuro-physiologically inspired, where the retinal cells, for the most part, transfer only the contrast information to the cortex, which at some stage performs reconstruction for perception. Their image reconstruction algorithm is based on least squares error minimization using gradient descent as well as its corresponding Bayesian framework for the underlying problem. The contrast map is computed using the Difference of Gaussians (DoG) operator at each iteration, which is then compared to the contrast map of the original image generating a contrast error map.

Their motivation of using contrast information is originated from the biological characteristics of retina. The main function of the primate retina, in doing spatial analysis, is to extract contrast information from the luminance distribution [26]. Two types of cells in retina, referred as on-center cell and off-center cell. The on-centre cells are activated when the centre of their receptive

fields are brighter than their surround and deactivated otherwise. The off-centre cells work the opposite way by turning on when the surround is brighter than the centre and off otherwise. Together these two cell types capture all the spatial information that is available in an image.

Their algorithm computes on-center and off-center contrast maps from the original image. If M represents the mask and I is the image, a contrast map is given by

$$C_m = M * I \qquad (2.19)$$

where * is the convolution operator and C_m is the composite contrast map combining values from both on and off-center contrast maps. This composite map without any additive noise is used in the algorithm for reconstruction. The following is the step by step algorithm for contrast based image reconstruction.

Algorithm 2.1 Image reconstruction from composite contrast map [26]

Step 1: *img in ← input*
Step2: *rf ← receptive_field_mask*
Step3: *contr d ← compute_image_contr(img_in, rf)*
Step4: *eta ← 0.8*
Step5: *img_out ←initial_value*
Step6: **while** *stopping_condition 6 is not true* **do**
Step7: **for all** *[x, y] in img_out* **do**
Step8: *contr_a ← compute pixel contr(x, y, rf)*
Step9: *contr_e ← cont_ d[x, y] – contr_a*
Step10: **if** *contr_e ≠ 0* **then**
Step11: *img_out[x, y] ← img out[x, y] + eta *contr_ e*
Step12: **if** *img_out[x, y] < 0* **then**
Step13: *img_out[x, y] ← 0*
Step14: **end if**
Step15: **if** *img_out[x, y] > 255* **then**
Step16: *img_out[x, y] ← 255*
Step17: **end if**
Step18: **end if**
Step19: **end for**
Step20: **end while**

Ma et al. [49] proposed a feasible and fast approach to attention area detection in images based on contrast analysis. They were able to generate a contrast based saliency map, compared to Itti's saliency map [22], and conduct local contrast analysis. Their contrast based saliency map is computed as follows:

An image with the size of M×N pixels can be regarded as a perceived field with M×N perception units. if each perception unit contains one pixel. The contrast value C_{ij} on a perception unit (i, j) is defined as follows:

$$C_{i,j} = \sum_{q \subset \Theta} d(p_{i,j}, q) \qquad (2.20)$$

where $p_{i,j}$ $(i \in [0, M], j \in [0, N])$ and q denote the stimulus perceived by perception units, such as color. $\boldsymbol{\Theta}$ is the neighborhood of perception unit (i, j). The size of $\boldsymbol{\Theta}$ controls the sensitivity of perception field. d is the difference between $p_{i,j}$ and q, which may be any suitable distance measure such as Euclidean distance or Gaussian distance according to applications. By normalizing to [0, 255], all contrasts $C_{i,j}$ on the perception units form a saliency map. The saliency map is a grey level image which the bright areas are considered as attended areas. Then a method referred to fuzzy growing is proposed to extract attended areas from the contrast based saliency map.

Figure 2.4 A vivid pencil sketch art work [19]

Contrast is the difference in visual properties that make an object (or its representation in an image) distinguishable from other objects and the background. Previous work that utilizes the contrast information of images has shown that contrast indeed is an important feature of an image to human visual attention system, and it can assist research in content based image analysis domain. In addition, the contrast information of objects is used by artists for pencil sketching, where the whole 3D world can be vividly depicted by the contrast among a set of grey levels on a 2D paper (see Figure 2.4). Therefore, we expect to adopt contrast information as one of the important features for the new image-based viewpoint metric, *Viewpoint Saliency*.

2.8. Template matching and segmentation

In a multi-camera system, template matching and image segmentation are important techniques for post processing data captured by multiple cameras. For instance, a fast and accurate template matching can help to recognize object from the images captured by different cameras. In this chapter, selected works on template matching and image segmentation are reviewed.

Omachi et al. [52] proposed a template matching algorithm, named as algebraic template matching. Given a template and an input image, algebraic template matching can calculates similarities between the template and the partial images of the input image for various widths and heights. In their algorithm, instead of using template image itself, a high-order polynomial decided by least square method is used to approximate the template image to match with the input image. Also this algorithm performs well when the width and height of the template image differ from the partial image to be matched.

Bong et al. [6] proposed a template matching algorithm for robot applications using grey level index table, which stores coordinates that have the same grey level, and image rank technique.

Their algorithm can find specific area under the given template query image with 30% Gaussian noise. They also presented a solution to object tracking using continuous query image tracking based on their template matching algorithm, which can compensate the situation when the system has different rotations or zooming levels for the object of interest.

Many have investigated into template matching that is invariant to certain changes of the template. For instance, Goshtasby and Ardeshir [17] presented a template algorithm in rotated images; Kim et al. [24] presented a rotation, scale, translation, brightness and contrast invariant grey-scale template matching algorithm originated from "brute force" solution, which performs a series of conventional template matching between the template and the input image by applying a series of changes, such as rotation, translation, etc to template image. However, their technique can substantially accelerate this process.

Additionally, Lowe [37] presented a method for image feature generation for objection recognition, referred as the Scale Invariant Feature Transform (SIFT). This approach transform an image to a large collection of local feature vectors, each of which is invariant to image translation, scaling, and rotation, and partially invariant to illumination changes and affine or 3D projection. The resulting feature vectors are called SIFT keys. The SIFT keys derived from an image are used in a nearest-neighbour approach to indexing to identify candidate object models. Collections of keys that agree on a potential model pose are first identified through a Hough transform hash table, and then through a least-squares fit to a final estimate of model parameters. When at least three keys agree on the model parameters with low residual, there is strong evidence for the presence of the object.

Previous image segmentation algorithms can be generally classified into two categories. One is feature-space based; the other is image-domain based. A Graph cuts based image segmentation method has recently attracted a lot of attention. Anjin et al. [2] proposed an automatic image

segmentation using mean shift analysis. It is demonstrated to be superior than previous graph-cuts based method on Berkeley segmentation dataset.

As it was mentioned previously, multi-camera views registration remain to be issues for multi-camera system. In order to acquire the best views of object(s) in a multi-camera system, we need to rely on previous works in template matching and segmentation domain for object recognition and segmentation.

2.9. Summary

Although image quality metrics link human attention with the assessment of image quality attributes such as sharpness or brightness, none of them address the problem of assessing the quality of viewpoints captured in images. Furthermore, for real time streaming applications on in cyber-physical environments, traditional 3D model based viewpoint selection algorithm cannot be applied because 3D model reconstruction is very difficult and time-consuming. Since response time is critical for QoS of real time applications such as VoIP and therefore influences their QoE [73], it is necessary to develop an efficient viewpoint selection and evaluation framework and an efficient cameras control scheme to enable real-time best view acquisition in cyber-physical environments.

3. Proposed Approach

3.1. Challenges and difficulties

Although a number of previous works have addressed the problem of best view selection, researchers have been making their efforts to develop metrics that can evaluate the viewpoint quality from different angles in three dimensional space. However, a lot of problems in the best view selection still remain unsolved, and furthermore, it is still challenging and difficult to provide a better solution to best view selection for real applications such as video conferencing and camera surveillance systems in cyber-physical environments. The following paragraphs seek to identify and analyze challenges of this work.

3.1.1. QoE versus QoS

Throughout years, people have never come to the consensus on the definition of best view(s). Some may argue that "the best view" can be individual dependent; however, it is always interesting to find out if there is any common sense among people that can shed light on evaluating the quality of different views for a given object.

To put this problem onto another level, we aim to improve the Quality of Experience (QoE) of users when intensive user interactions are involved in the multimedia environments. Previously, many successes have gained on studying Quality of Service (QoS), which really makes us consider a series of questions such as "What is quality of experience, what are the relationship between QoE and QoS, and how can QoE be improved based on the efforts made on QoS?" Wanmn et al [73] proposed a theoretical framework of modeling QoE. In their work, the

relationship between QoS and QoE is addressed as a causal chain of "environmental influences → cognitive perceptions → behavioral consequences". In order to solve the problem of best view selection from an angle that maximizes users' quality of experience, a thorough understanding between human perception and cognition is required to narrow the semantic gap (i.e., the differences between human activities, observations and computational representation).

3.1.2. Two dimension versus three dimension

Traditional solutions largely rely on the availability of 3D models, which are difficult to construct in real time. In previous works, best view(s) selection generally requires prior knowledge of the geometry of the scene or objects and relies on the availability of the 3D model of them. Selections are usually made assuming that all the possible views can be captured by cameras. This is useful in a completely synthetic computer graphics environment but it is not applicable to cyber-physical environments which consist of fixed number of sensors and require real-time processing. Alternatively, we are trying to develop a new image-based measurement of viewpoint qualities, named as viewpoint saliency (VS). We hope to base our viewpoint quality metric on two dimensional information, i.e., features extracted from images of interested object(s), and reduce the computation complexity caused by 3D model reconstruction.

3.1.3. Online versus offline

Real applications call for real time processing and online response. Apart from 3D model reconstruction, traditional best view selection algorithms are hampered by the large amount of

computation overhead. And none of them can guarantee QoS such as timely response to users' request, which can be detrimental in real applications such as video conferencing or camera surveillance systems. In our approach, we hope to first develop 2D based metric, and then control multiple cameras to select best view of objects and make sure the best results can be returned to users in real time.

Additionally, other problems are such as: "if only limited number of cameras are available and given their positions are fixed, what if none of the cameras can capture a good view of the object(s) with its limited strength (pan, tilt, and zoom)?" and traditional issues such as object recognition and segmentation in multi-camera systems.

As it is stated above, a number of problems remain to be solved. Therefore, we shall try our best effort to provide them with solutions.

3.2. Motivation and background

Remote monitoring and control mechanisms have long been desired for use in inhospitable environments such as radiation sites, under-sea and space exploration [18]. Traditional remote monitoring systems generally consider user's selections as the region of interest, instead of object of interest. Computation is generally performed based on regions, instead of semantic objects. However, in applications such as video conferencing, the concept of objects (e.g. the remote person) is important to users. Therefore, we would like to develop a metric to evaluate the viewpoint quality of specific objects. This viewpoint evaluation metric should be computable in real time without reconstructing the 3D model of the object. Previous research in visual attention analysis and image quality assessment provides ideas of 2D feature assessment. We can combine

it with ideas from 3D viewpoint selection to develop a new 2D based viewpoint evaluation metric. This reduces the computation cost since it works entirely in the 2D space.

3.3. Image based viewpoint quality metric

3.3.1. Viewpoint saliency (VS)

Our proposed viewpoint saliency (VS) metric is able to compute potential scores (ranging from 0 to 1) for the quality of various viewpoints of objects captured by images. The definition of VS is as follows:

Let F = **{**F_1, F_2, F_3, F_4……**}** be the set of features extracted from an image (a view) of an object, let P = {p_1, p_2, p_3, p_4...} be the set of descriptors that describe the features included in F, where $p_i \in R$ *and* $p_i \in [0,1]$. And every single feature in F has one (or more than one) descriptor (s) in P. The relative importance of each descriptor in form the set of weights W = {w_1, w_2, w_3, w_4...}, where $\sum_i w_i = 1$. Let *VS* be the score for the quality of this view, i.e. viewpoint saliency (VS)

$$VS = \sum_i w_i p_i \qquad (3.1)$$

Descriptors are real numbers ranging from 0 to 1 to interpret the strengths of features shown in the tested images (views). The larger the value of a descriptor, the more information is conveyed by its associated feature. So far, we have found two features are important to the quality of a given viewpoint, one is the contrast level within object region, denoted as p_c, the other is the projected area of the object, denoted as p_a. And initially, we assume they are of same importance, hence, $w_1 = w_2 = 0.5$. Viewpoint saliency (VS) is computed as:

$$VS = w_1 \times p_c + w_2 \times p_a \qquad (3.2)$$

In the following paragraphs, we will give detailed explanations of the two descriptors, i.e., p_c and p_a and the possible further extensions of the above definition.

Contrast level descriptor p_c

Given an image I, where its region of interest is the region that contains an interested object, referred as object region, the contrast level descriptor p_c of the object region is computed as:

$$p_c = \frac{1}{N_p} \sum_{p_{i,j} \subset O} C_{p_{ij}} \qquad (3.3)$$

where O indicates a bounded object region, $p_{i,j}$ indicates one perception unit within the object region. One perception unit can either be a single pixel or a sub-region of O, which decides the granularity of p_c. N_p is the total number of perception units within the object region O. $C_{p_{ij}}$ is the contrast level value of the perception unit $p_{i,j}$ obtained from the contrast map of I.

The contrast map of an image is a map in which each perception unit is encoded with a contrast level value compared with its neighborhood. The idea of constructing contrast map is from previous work on image construction from contrast information [26]. The calculation of the contrast map is based on the contrast-based visual attention model [49] (see section 2.7.2), which is proved to be capable of obtaining equally effective results with Itti's visual attention model [22] yet has less complexity and requires less computation time. Examples of a contrast map of a general object and a contrast map of human face are shown in Figure 3.1.

a **b**

Figure 3.1 Original images and their contrast maps
a. General object; b. Human face

The contrast maps shown in Figure 3.1 (a) and (b) are computed under the stimulus of color, and in the map, the contrast of a region can be visualized as the brightness of the region. The brighter the area, the higher contrast it is perceived. The following paragraphs will give a detailed explanation of computing the contrast map.

The method of constructing the contrast map for a given image is as follows:

An image with the size of M×N pixels can be considered as a perceived field with M×N perception units if each perception unit contains one pixel. The contrast value $C_{p_{ij}}$ on a perceived pixel at location (i, j) of the image is defined as follows:

$$C_{p_{ij}} = \sum_{q_{m,n} \subset \Theta} d(p_{i,j}, q_{m,n}, stimulus) \qquad (3.4)$$

where $p_{i,j}$ $(i \in [0, M], j \in [0, N])$ denotes a single perception unit, and $q_{m,n}$ denotes one neighborhood perception unit surround $p_{i,j}$. Θ is the set of all the neighborhood perception units of $p_{i,j}$. Notice that the size of Θ controls the sensitivity of perception field: the smaller the size of Θ is, the more sensitive the perceive field is. For instance, Θ can be a 3×3 neighborhood square window around $p_{i,j}$, this will yield 8 neighborhoods. $stimulus$ denotes the stimulus of the contrast among perception units, for instance, it can be color, texture, or orientation, etc. $d(p_{i,j}, q_{m,n}, stimulus)$ measures the difference between $p_{i,j}$ and $q_{m,n}$ under a certain stimulus such as color, which may employ any suitable distance measure such as Euclidean distance or Gaussian distance. In the experiment introduced in section 3.4, Gaussian distance is used. By normalizing to [0, 1], all contrast values $C_{p_{ij}}$ of the perception units in the perceived field (i.e., the image) form a "contrast map" that stores the contrast value for each perception unit, (i.e. each pixel).

Currently, in our experiment, color is proved to be a good stimulus for computing contrast map. In Figure 3.1 contrast maps are computed under the stimulus of color in LUV space, especially, U

and V components in LUV space are used to compute the distance between one perception unit and its neighborhoods, the following distance measure is used:

$$distance = a(1 - e^{-\frac{d}{2q^2}}) \qquad (3.5)$$

where *a* and *q* are constants, *d* is Euclidean distance in 2D space.

To reduce the number of colors in the image, a color quantization algorithm is applied before calculating the contrast map. For the neighborhood window size, 3 pixels by 3 pixels square window is used.

Projected area descriptor p_a

Projected area is used as important information in the theory of viewpoint entropy (see section 2.2.1). Without any prior knowledge of three dimensional structure of interested object, project area can be a good descriptor for interpreting the quality of viewpoint in two dimensional space.

Given an image of an object with a rectangular object region, the projected area descriptor is computed as:

$$p_a = \frac{aWH}{MN} \qquad (3.6)$$

where *W* and *H* are width and height of the object region. M and N are height and width of the image. a is the scaling number.

As it is mentioned, both p_c and p_a range between 0 and 1, and describe the amount of information conveyed by contrast level and projected area in the objects' images. Then, substitute p_c and p_a in formula (3.2) with formula (3.3), formula (3.4) and formula (3.6), we obtain

$$VS = \frac{w_1}{N_p} \sum_{p_{i,j} \subset O} \sum_{q_{m,n} \subset \Theta} d(p_{i,j}, q_{m,n}) + w_2 \frac{aWH}{MN} \qquad (4.7)$$

where w_1 and w_2 are the weights of p_c and p_a, indicating their relative importance. Initially, $w_1 = w_2 = 0.5$.

Flexibility and extensibility of the definition

The above definition of viewpoint saliency (VS) is extensible and flexible. Later on, various aspects can be improved through in-depth research without affecting the structure of the formulation. Aspects to be improved can be summarized as follows:

(1) More features can be researched on the viewpoint quality evaluation, and they can be easily incorporated into the formulation by developing specific descriptors for the features.

(2) The method of obtaining descriptors can be improved without changing the formulation.

(3) Relevance feedback can be included to improve the evaluation result. It can be done by online asking users to re-order the results (from good to bad) generated by initial best view selection by their preference, and based on users' feedback, the relative importance of each feature descriptor can be adjusted through re-weighting the descriptors in the viewpoint saliency (VS) metric. This idea may provide an opportunity to improve best view selection to another level: to make the selection adaptable to individuals' preferences. For different users, the importance of one features to the viewpoint quality may be different, we can record this and apply different weight parameters for different users.

In addition, when applying VS metric, we assume that the scale does not change for each of camera view, which means that zoom parameters of cameras are not fully utilized at this stage. However, this can be improved in the future by considering the object region size versus the aspect ratio and size of a camera view in the VS definition.

3.4. Experiments

The image based viewpoint quality measure, viewpoint saliency (VS), can eliminate the time and reduce computation cost required for 3D model reconstruction of objects; and hence, is more desirable for real time applications. The experiments below we have conducted seek to study the effectiveness and utility of the of our proposed viewpoint quality metric VS.

3.4.1. Methods

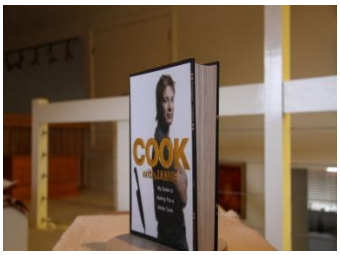
In order to test our proposed metric VS, we first conduct tests on the contrast level descriptor p_c, on images of general static objects taken from different viewpoints. Then in order to make our approach feasible for conferencing applications via Internet, we conduct tests on different human views.

General Objects

First, four objects of different size, color and texture are selected: a book, a laptop, a porcelain statue and a toy car. Their images are shown in Figure 3.2. A rotating table was used to take images at 30 degree interval over 360 degrees resulting in 12 images for every object. The lighting conditions and objects' scale for every view of each object is kept the same.

a

b



c

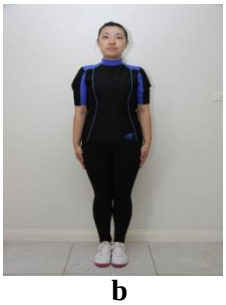
d

Figure 3.2 Images of selected general objects
a. Book b. Laptop c. Porcelain statue d. Toy car

Humans

Usually, when people do conferencing via live streaming applications such as Skype, or Google video chat, humans are the principle objects of interest. We thus want to obtain the best possible views of people through the available cameras using the same 2d based view evaluation metric, VS. For humans, three different positions (Figure 3.3) are considered: full body sitting, full body standing and human face. For each position, we start from the view angle where the frontal face is shown and mark it as the image at zero degree. Again we take images at 30 degrees interval to obtain 12 images covering the full circle of viewpoints for each of the three positions of Figure 3.3; and uniform lighting and the same object scale are maintained.

a



b

c

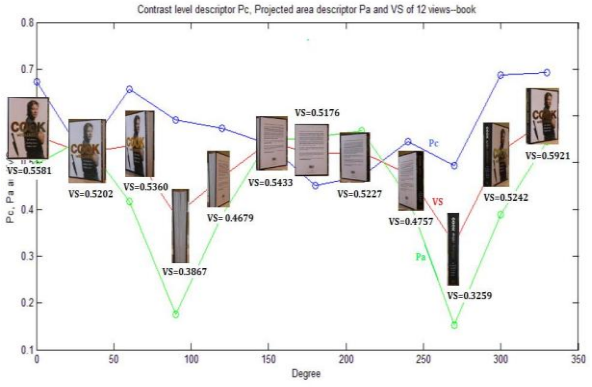
Figure 3.3 Images of humans with different positions
a. Sitting human b. Standing human c. Human face

3.4.2. Results

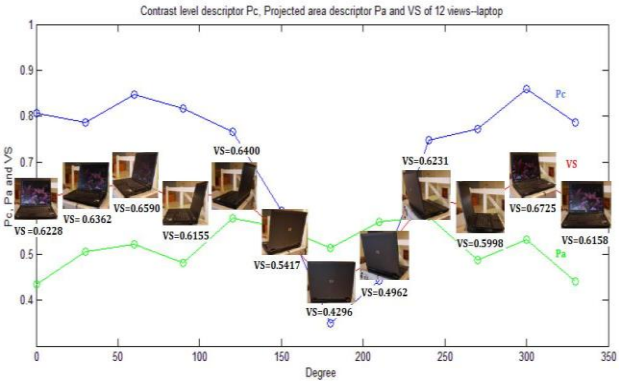
After the images are obtained, the contrast level descriptor $p_{c,}$ projected area descriptor p_a and VS of all images of selected objects and humans are computed. The experiment results of selected general objects shown in Figure 3.2 and human objects shown in Figure 3.3 are presented in Figure 3.4 and Figure 3.5. In Figure 3.4 and Figure 3.5, 12 views of objects are arranged according to their computed VS score, and the contrast level descriptor p_c (blue line) and projected area descriptor p_a (green line) of each 12 views are also plotted in the same graph.

3.4.3. Analysis

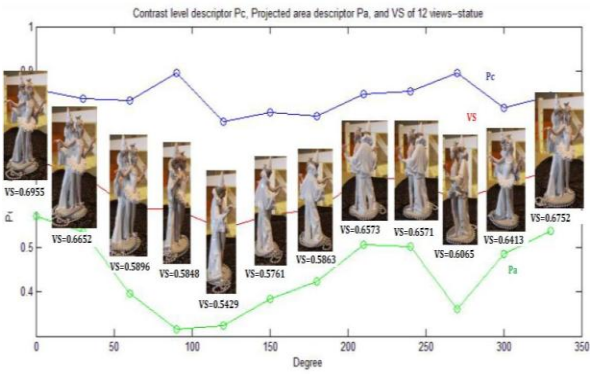
From Figure 3.4 and Figure 3.5, we can see that VS indeed provides us a fast and effective alternative to evaluate viewpoint quality in 2d space. High VS score indicates the object view is "good" and low VS score indicates the view is not so good. Additionally, from Figure 3.4, we can see that VS not only is able to deal with general static objects, human being, as a matter of fact, can be handled by VS as well.

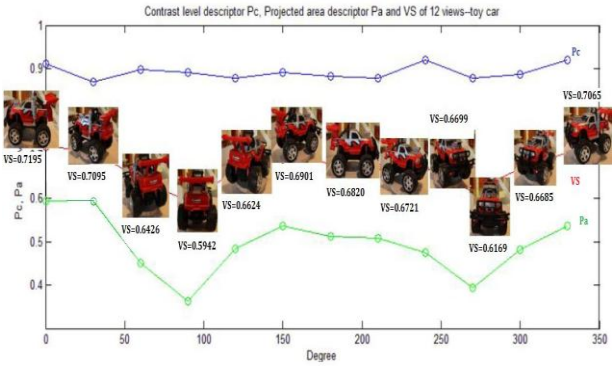


a



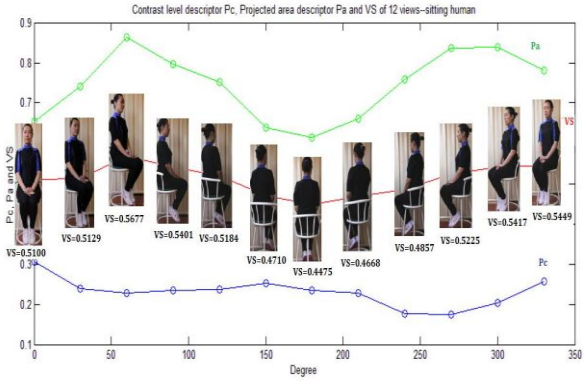
b

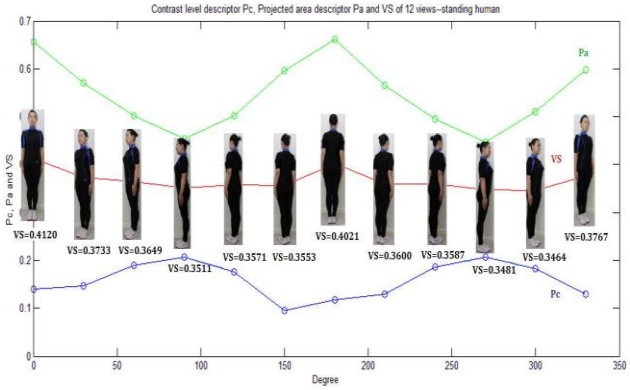




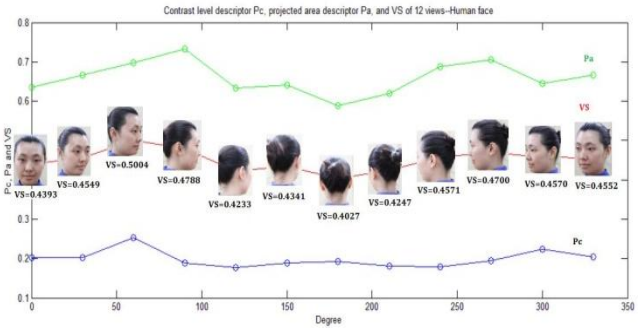
c **d**

Figure 3.4 12 views of general objects ranked by their VS scores
a. ordered 12 views of a book; b. ordered 12 views of a laptop;
c. ordered 12 views of porcelain statue;
d. ordered 12 views of a toy car





a **b**



c

Figure 3.5 12 views of human objects ranked by their VS scores
a. 12 views of sitting human; b. 12 views of standing human; c. 12 views of human face

From above experiments we have found that two factors can greatly affect the computational result of VS, they are (1) the strong texture of objects; (2) the changes in lighting condition. However, in our current work, we assume that the lighting condition remains the same for one

cycle of the best view selection (when multiple cameras from different viewpoints are simultaneously capturing images of the object of interest). Therefore, in Figure 3.5(c), we can see that the back view containing purely hair of a human was not evaluated to be a good view. Additionally, for evaluating the viewpoint quality of humans, we found that the projected area descriptor P_a could be very deceiving especially because of the shapes of humans' bodies have tremendous variability. For instance, the projected area of someone's full body standing side view could be significantly larger than the frontal view; however, it is not the case for others.

Notice that in Figure 3.5(b), the back view of the human body is evaluated to be almost as good as frontal view, mainly because of its larger projected area (see projected area descriptor P_a plotted with green line); however it does not possess much contrast and present much information to us. We are still working on improving this drawback of VS for evaluating views of humans, which remains to be an interesting challenge.

In order to compare the results of VS with a 3D viewpoint measure - Viewpoint Entropy (VE), and with actual users' choices of best views, we conducted the comparison based on 12 views of the book (Figure 3.4(a)) and 12 views of the laptop (Figure 3.4(b)).

The comparison results are shown in Figure 3.6, the ranking are from 1 to 12, where 1 indicates the best view and 12 indicates the worst view. For viewpoint entropy, we first compute orthogonal frustum entropy [68] for all the 2D images, and then we rank the 12 views according to the computed entropy (plotted in green in Figure 3.6).

In order to minimize the drawback of viewpoint entropy algorithm, i.e. discretization instability (stated in section 2), we manually segment the book and the laptop into roughly 2 faces. Examples of face segmentations are shown in the upper right corner in the plotted graph in Figure 3.6. For user study, 13 users were invited to a test for ranking 12 views of the above objects (see Figure 3.2 and Figure 3.3) from good to bad based on their own perception. Users were first told about the aim of our research, i.e., best view selection of object(s) and then the following question was asked together with showing 7 groups of 12 images of 7 tested objects:

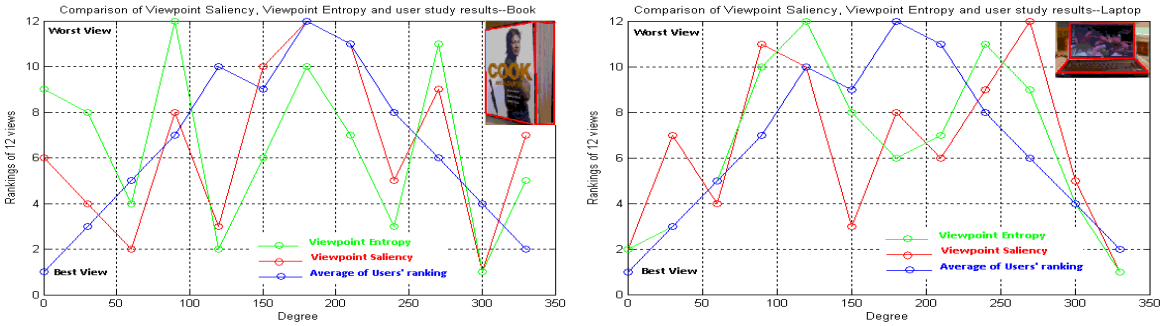
"Please rank the 12 views of the following objects from good to bad based on the amount of information they present according to your own perception."

Then, groups consisting of 12 Views of general objects including the book, the laptop, the porcelain statue, and the toy car (see Figure 3.2) as well as human being including standing human, sitting human and human face (see Figure 3.3) were shown to the users one after another. In the test, users were not allowed to communicate with each other for the answers.

After the test, we collected their responses and computed the average users' rankings for the 12 views (plotted in blue in Figure 3.6). The red dotted line in Figure 3.6 indicates rankings provided by our proposed metric, Viewpoint Saliency.

We also compute the correlation between Viewpoint Saliency (VS), Viewpoint Entropy (VE) and users' rankings on 12 views of other general objects shown in Figure 3.2 and human objects shown in Figure 3.3. Table 3.1 shows the correlation results between VS, VE, and users' ranking. Note that computing VE requires segmentation of an object's faces, but it is difficult to define the concept of "faces" and conduct segmentation for complex objects such as porcelain statue (Figure 3.2(c)), toy car (Figure 3.2 (d)), and humans (Figure 3.3); hence VE is not able to provide evaluation for complex objects (indicated as "N/A" in Table 3.1). We therefore only compare the ranking results obtained from our proposed metric VS with users' responses for complex objects and humans.

From Table 3.1, we can see that for complex objects such as porcelain statue and toy car, results generated by VS have strongest correlation with user's perception. However, for the book and the laptop, results generated by VS have strong correlation with results generated by VE, yet less correlation with user's ranking. This probably because some users may consider the back views of the book and the laptop to be good as well, as they provide other knowledge such as context or brand information, which is not considered by VS or VE. As for the human objects, VS has less satisfying results (especially for standing human views), which once again indicates this interesting challenge for us to improve our VS metric for evaluating human views.



a **b**

Figure 3.6 Comparison of Viewpoint Saliency, Viewpoint Entropy and users' ranking a. 12 views of book b. 12 views of laptop

Table 3.1 Correlations between 12 views ranked by Viewpoint Saliency (VS), View Entropy (VE) and users' ranking

General Objects	***Book views***	***Laptop views***	***Statue Views***	***Toy-car views***
corr(VS, VE)	0.6713	0.7692	N/A	N/A
corr(VS, usr)	0.3427	0.4755	0.8601	0.8182
corr(VE, usr)	0.3357	0.6993	N/A	N/A
Human Objects	***Sitting Views***	***Standing Views***	***Face Views***	
corr(VS, usr)	0.6434	0.2657	0.5245	

In conclusion, compared to the existing measure, i.e., viewpoint entropy, our proposed metric, viewpoint saliency can eliminate the segmentation of "faces" of objects and provide reasonably good viewpoint evaluation results for both simple and complex objects, even for humans. Its generality for all types of objects and simplicity in computation facilitates best view acquisition for real time applications in cyber physical environments.

3.5. Real time best views selection as energy minimization

Our proposed metric VS eliminates the cost required for 3D model reconstruction and enables best view selection and acquisition to be achieved in real time. We first propose an approach of doing 2d best view selection by mapping the 3d viewpoints of an object to its 2d images, this

approach is illustrated in Figure 3.7. As shown in Figure 3.7, all the viewpoints of a given object form a sphere around the object, and at each viewpoint, an image can be taken to represent the view of the object. In reality, we assume that there are a few cameras around an object so there are only limited numbers of viewpoints that can be captured by multiple cameras. Therefore, we can down-sample the sphere by only taking the camera-reachable viewpoints and then flatten it out to form a graph, where each node represents an image of a possible viewpoint that can be reached and captured by cameras. The edge in the graph represents the relationship between different views in terms of camera movements (for instance, pan or tilt).

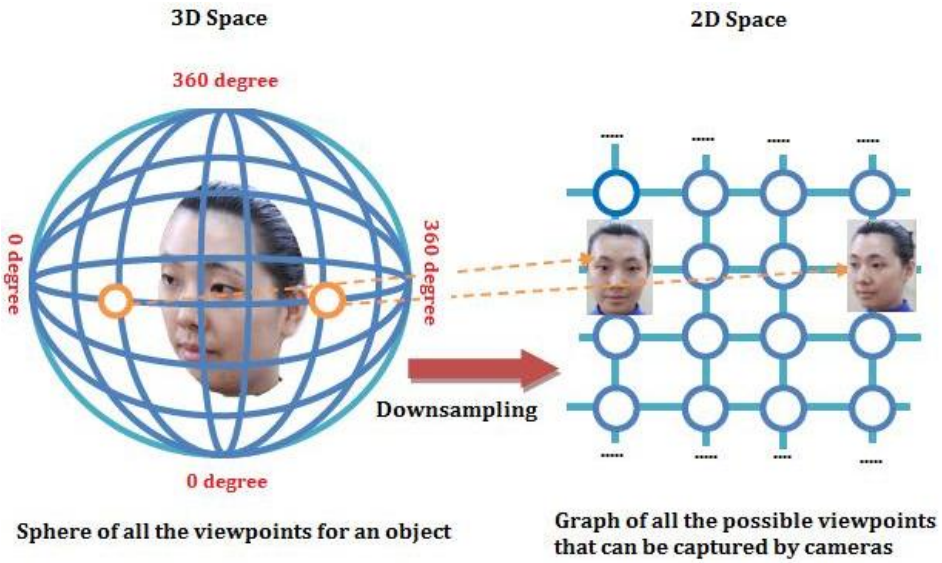


Figure 3.7 Mapping from 3d space to 2d space

After performing the above mapping, real time best view selection and acquisition can be done in the 2D space (i.e. the graph) without constructing the 3d model of the object and the best view selection problem can be transformed into a "Best Quality Least Effort" task, where we want to obtain the best possible view(s) of interested object(s) with the least amount of cost. "Quality" and "Cost" are two important facets which can be traded-off against each other. They can be formalized as two terms of an energy function, and 2D best view search can be modeled as a

finite state transition problem for energy minimization. Our proposed energy function is defined in the following paragraphs.

3.5.1. Proposed energy function

Assume a remote environment containing N Pan-Tilt-Zoom (PTZ) cameras, for a given object O, the set of views (images) of O can be captured by these N cameras is denoted as V = $\{v_1, v_2, \ldots, v_m\}$ (m >>N). All the elements of V forms a undirected graph ***G*** (V, E) with m nodes, the edges of the graph is defined by E=$\{e_1, e_2, \ldots, e_k\}$, which also indicates the relationships (in terms of predefined one step camera movement such as left pan 10 degree or up tilt 10 degree) between individual views in V. For instance, if one camera is at position where v_i can be captured, and the camera needs to proceed one step (e.g., 20 degree) of movement (e.g., pan) to obtain v_j, then there is an edge e_{ij} between v_i and v_j. If v_i cannot be transformed to v_j through one step camera movement, there is no edge between the two nodes. Each edge in E is associated with a triple, i.e.,

$$e_i = (u_i, t_i, v_i) \tag{3.8}$$

where the triple indicates the start node u_i, end node v_i and the time required t_i for moving the camera from u_i to v_i.

S = {S_0, S_1,.., S_t}denotes the states of N cameras throughout the process of best view selection, the transition from one state to another in **S** is made by one step of N cameras' movements. Each element of **S** is a subset of V and contains the current views monitored by N cameras, i.e., $S_i = \{v_{ij}\}_{j=1}^{N}$. S_0 is the initial cameras' states; S_t is the final cameras' states after selection, i.e., the best possible views captured by N cameras.

The energy of one cameras state S_i ($S_i \in \boldsymbol{S}$) takes both "Quality" and "Cost" into consideration. To achieve quality maximization and cost minimization, the image based best view selection and

acquisition is formulated as finding the N camera state S_i in the undirected Graph $\boldsymbol{G}$ (V, E) formed by view set V where the energy of S_i, i.e., $\mathbf{E}$ (S_i) is minimized.

$$\boldsymbol{E}(S_i) = \alpha_1 \boldsymbol{E}_{quality}(S_i) + \alpha_2 \boldsymbol{E}_{cost}(S_i) \tag{3.9}$$

$$\text{s.t.} \quad S_i = \{v_{ij}\}_{j=1}^{N}$$

$$S_i \cap V \neq \emptyset$$

where $\boldsymbol{E}_{quality}(S_i)$ and $\boldsymbol{E}_{cost}(S_i)$ denote the "quality" energy term and cost energy term of state S_i , α_1 and α_2 are predefined weights balancing the strength of each energy term. Initially, $E(S_0) = \infty$.

The analogy of above cameras' state transition driven by minimizing energy is illustrated in Figure 3.8, where the camera state transitions from S_m to S_n by one step of cameras' movements under the condition that the energy of S_n is lower than that of S_m. In Figure 3.8, the edges between the start nodes *u* and end nodes *v* are denoted as *e*.

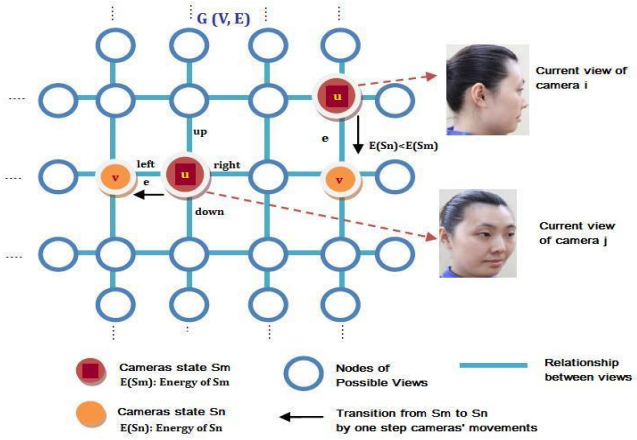


Figure 3.8 Cameras' states transition driven by minimizing energy

3.5.2. The "Quality" term

The quality energy term measures the quality of selection, i.e. the quality of selected viewpoint for a given object. In Section 3, we introduced our 2D based measure VS which provides a mean to evaluate the viewpoint quality. Therefore, we define the "Quality" aspect of a cameras state by measuring its total improved viewpoint quality from initial state S_0 for each camera view.

$$\boldsymbol{E}_{quality}\ (S_i) = \sum_{j=1}^{N}\{\boldsymbol{VS}(v_{0j}) - \boldsymbol{VS}(v_{ij})\}^2 \qquad (3.10)$$

where $S_0 = \{v_{0j}\}_{j=1}^{N}$ is the initial camera state, and its elements are initial views of a given object captured by each cameras as images, ***VS*** is the image based viewpoint quality metric introduced in section 3.3. N is the camera number, $S_i = \{v_{ij}\}_{j=1}^{N}$ is the cameras state, which contains N current views of the object.

3.5.3. The "Cost" term

The cost of the best view selection is a very important factor for real time application as it directly affects the response time of our result.

Generally, the cost of time is incurred by two factors, i.e., computation and cameras movement. In our experiments, we found that it typically takes only 100~300 msecs to compute the VS out of a 4CIF (640× 480) camera view. Since VS is computationally inexpensive, we can neglect it.

Therefore, the cost is mainly due to camera movement. The cost energy term is determined by measuring the total amount of cameras' movement required for the transition of cameras states from the initial state S_0 to the current state S_i.

$$\boldsymbol{E}_{cost}\ (S_i) = \sum_{k=0}^{i}\sum_{j=1}^{N} t_{kj}^{\ 2} \qquad (3.11)$$

where t_{kj} is the associated time variable (i.e., the time consumed by one step camera moving from one view to another) for an edge in Graph G (see section 3.5.1 formula (3.8)), which links one camera state with another. N is the total number of cameras.

3.5.4. Cameras control

In this thesis, the type of sensor we consider is the PTZ (Pan-Tilt-Zoom) camera. Assume that there are N numbers of cameras in the remote environment. We want to move all the cameras from initial state S_0 to the final state S_t which has minimum energy. In order to control multiple cameras for solving the energy minimization, we apply the idea of multi-scale search. For each camera, initially, the search area is formed by all the possible camera positions for viewing the object of interest. Then, each camera performs a full pan and tilt within the search area, taking images along the vertical and horizontal axis of the search area, and these images of viewpoints can serve as samples for predicting the energy of sub area and guiding cameras move towards the predicted energy minimizing sub area, which is ¼ of the original search area for the next scale of search. We set the stopping criteria as the search area is smaller than one step of camera movement (20 degree pan/tilt). Finally, when all the cameras have stopped moving, they are at the cameras state with the minimized energy, and N best possible views of a selected object given by N cameras are obtained. The final best view is then selected by comparing the VS score of final views of each camera. This idea is illustrated in Figure 3.9. Notice that the initial search area for each camera depends on the pan/ tilt parameters of the camera. And the total energy of a cameras state is computed for all the cameras in the system (i.e., there is one single graph shown in Figure 3.8, however each camera performs a multi-scale search shown in Figure 3.9).

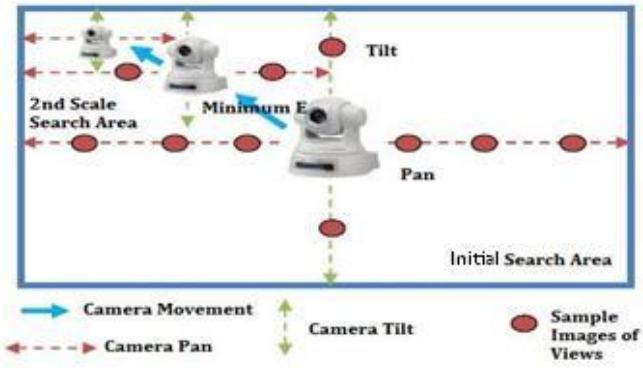


Figure 3.9 Multi-scale search of a single camera

Based on above camera control scheme, we present our algorithm for obtaining the best view of user selected object(s) through energy minimization driven cameras state transition.

Our proposed algorithm of camera control for real time best view selection and acquisition is as follows:

Algorithm 3.1: **Real Time Best View Selection and Acquisition**

Input: Initial state of N cameras $S_0 = \{v_{0j}\}_{j=1}^{N}$
Output: Final state of N cameras $S_t = \{v_{tj}\}_{j=1}^{N}$
and the best view v_m $(v_m \in S_t)$
Initialization: x = 180 (degree)
OneStep_move = 20(degree)
S = $\{S_0\}$
Search_areas = $\{A_j\}_{j=1}^{N}$
While (x > = OneStep_move**)**
For each of the N cameras
Step 1: Based on user selected the rectangular region, match it with the current view of the camera using scale invariant template matching method (SIFT) [37]
Step 2: Do a full pan and tilt within its search area A_j, taking images at sample positions (every x degrees along the vertical and horizontal axis) (See illustration in section 3.5.4. and Figure 3.9)
Step 3: Compute the energy **E** (improved quality versus camera movements) of all possible camera state with sample images taken by all the cameras. (See formula (3.9))
Step 4: Based on computation in **Step3**, predict the next scale search area A_j', which could yield to a total energy minimization. (See section 3.5.4., Figure 3.9)
Step 5: Move the camera to the center of the new search area
Step 6: x = x/2
$A_j = A_j'$
S = S + $\{S_i\}$
End For

End While
$S_t = S_i$
For each of the N final camera views
Compute their VS scores (see section 3.3, formula (3.7))
End For
Select the camera view with the highest VS score as the best and yield the control of this camera to users

---

4. System and Experimental Results

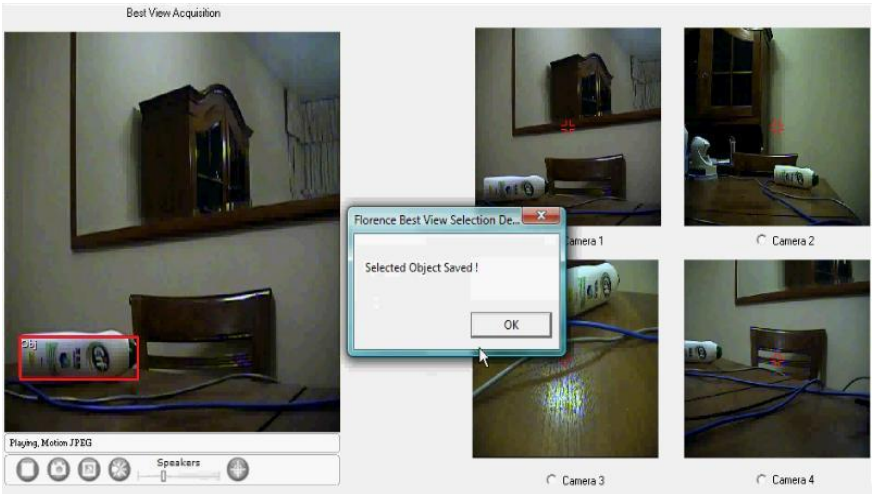
We implemented our algorithm with four Axis 214 PTZ network cameras on VC++ platform. Each camera can be connected to a local network or internet, and our program can facilitate users to remotely operate multiple cameras or to automatically acquire the best view of selected objects, including humans. This is useful in real time communication applications via Internet, such as video conferencing via VoIP. Additionally, we describe our extended features for WWW based cameras control and best view acquisition, which can allow low cost public access to remote observation, navigation and education systems. Readers are welcome to watch the video demonstration of our system online at http://www.youtube.com/watch?v=gTIvg3eoAjM

4.1. The user interface

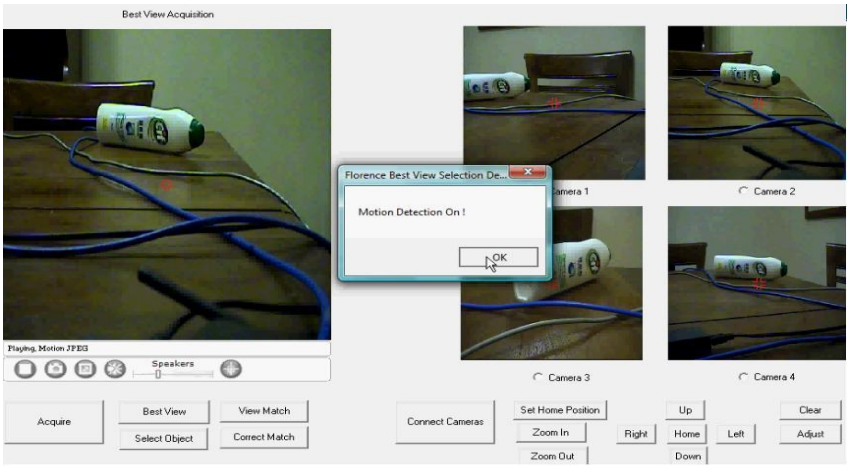
The user interface is shown in Figure 4.1(b). On the right side, four small views are current views of connected four different cameras, users can manipulate the cameras (pan/tilt/zoom) by clicking either on the screen monitor or the buttons (home / up /right / left / down) below. On the left side, the best view will be finally presented in the large screen monitor after computation. Initially, the large screen monitor shows the view of camera one; but users have the flexibility to switch the views of every other cameras onto the large screen by clicking the radio button below them. And users can choose to obtain the best view(s) of object(s) of interest through our automatic control of cameras or through their manual operation of cameras.

4.2. Best view acquisition of single object

Our real time best view acquisition result of a single object is demonstrated in Figure 4.1. In Figure 4.1(a), a white bottle is selected, and after camera adjustment, the best view is given by camera 2 in Figure 4.1(b). In our system, after the first best view acquisition, a motion detection mechanism is switched on, and it will be able to detect the motion of the object of interest and make adjustment periodically (every 90 frames). A detailed motion detection scenario is illustrated in section 4.4. More video result can be found at http://www.youtube.com/watch?v=nSm8wiCJlEQ



a

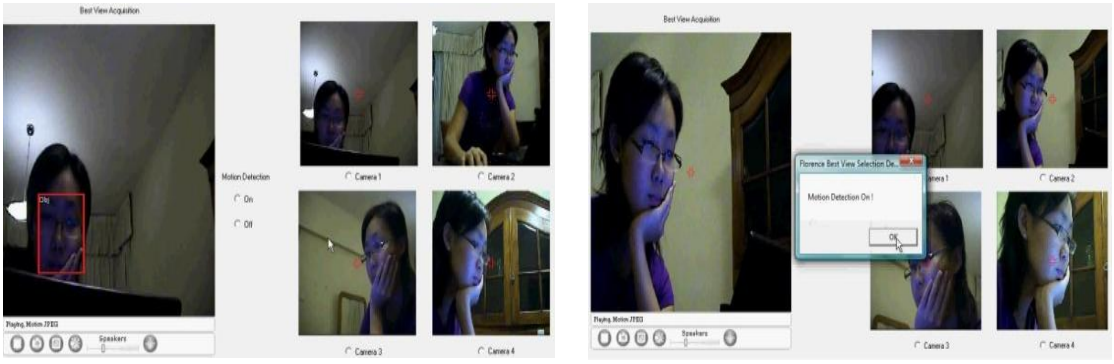


b

Figure 4.1 Best view acquisition of single object
a.User selection; b. Acquisition result

4.3. Best view acquisition of human

The best view acquisition result of human is shown in Figure 4.2. In Figure 4.2(a), a human face is selected, in Figure 4.2(b), final results are shown. The video clip of this result is available at http://www.youtube.com/watch?v=WpiTgCHoXqI



a **b**

Figure 4.2 Best view acquisition of human
a. User selection; b. Acquisition result

4.4. Extensions for web-based real time applications

With Web 2.0, web based applications have become more and more popular due to their low cost and less complexity and are offered by more and more vendors such as Google and Amazon. WWW based cameras control can be handled by sending HTTP request to network cameras using GET/POST method with associated pan/tilt/zoom parameters. The web-based cameras control interface shown in Figure 4.3 below is implemented in JavaScript, and our VC++ based best view computation (mainly for template matching and Viewpoint Saliency computation) can be built as Win32 DLL for guiding WWW based cameras control through JavaScript. In this

chapter, we especially demonstrate the extended features of our system for WWW based applications.

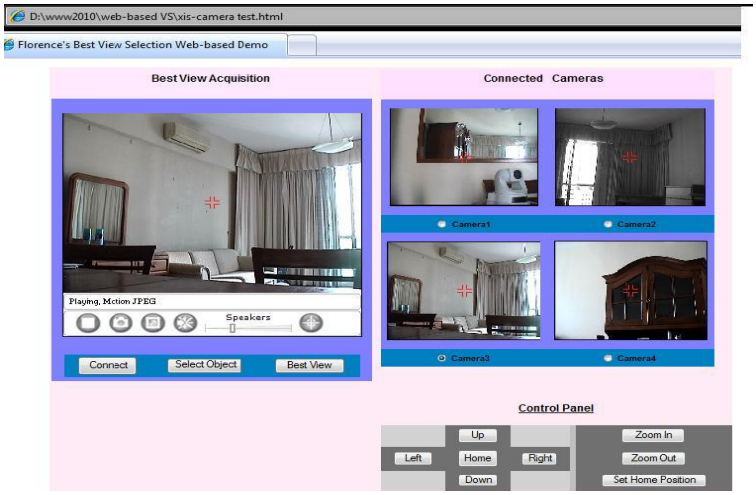


Figure 4.3 Remote Monitoring and Tele-operation of Multiple IP Cameras via the WWW

Sometimes, users may like to view more than just one object at one time, and our system is able to provide best view acquisition of multiple objects. We handle this by individually computing the Viewpoint Saliency (VS) of multiple object regions and add them together as the overall VS of the captured view. Figure 4.4 shows the results of two selected objects: one is a candle holder (red), the other is a sticky tape (yellow).

a **b**

Figure 4.4 Best view acquisition for Multiple objects a.User selection; b. Acquisition result

In the application of video conferencing, it is useful to constantly provide users the best view of objects of interest (e.g. humans) based on their first time selection. Sometimes, the object of interest may move to a new position, and our system is able to detect the motion by storing the acquired best view of the object as a reference frame and periodically compare the difference between the current view and the reference; if the difference is larger than a given threshold, we trigger the event that the object has moved and accordingly re-adjust all the cameras. Figure 4.5 demonstrates this scenario. In Figure 4.5(a), the yellow tape is selected by user, and in Figure 4.5 (b) the first–time best view acquisition result is presented. Then, we move the position of the tape, our system detected the move (Figure 4.5 (c)) and made the adaptive adjustment. In Figure 4.5(d), the re-adjusted result is shown. However, the motion we assume here is only slow motion, and we are still working on improving this mechanism. In the future, we hope to continuously provide the best views of object(s) of interest.

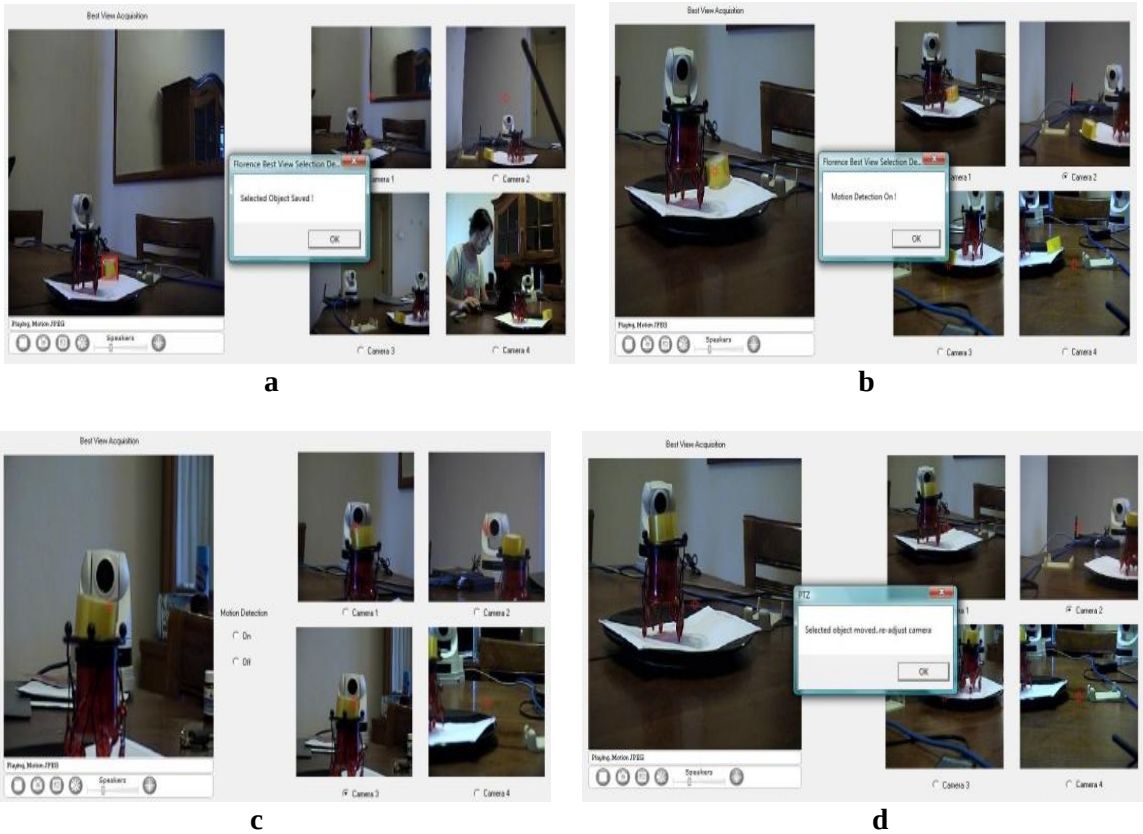


Figure 4.5 Best view acquisition for object with motion
a.User selection; b. Acquisition result c. Move of object d. Re-adjustment

4.5. Quality of Experience (QoE) evaluation

Due to the subjectivity of QoE, there is few standard quantitative metric for evaluating the QoE of a multimedia system. Previous work [73] has addressed the QoE construct, QoS construct and their correlations in user experience modeling. In order to evaluate the QoE of our system, based on the correlations between QoS construct and QoE construct [73], we adopted two important criterion in QoS construct: "*interactivity*" and "*subjective consistency*", and interpret them as the "*the degree of interactivity to satisfy users' needs*" and "*the level of consistency to user desired results*" and all the representative dimensions in QoE construct to evaluate the QoE of our system. Five representative dimensions in QoE construct [73], namely, concentration, enjoyment, tele-presence, perceived usefulness and perceived ease of use are summarized and interpreted as another three criteria in our evaluation: "*Ease of use: the level of ease to operate and use the system to achieve user desired results.*", "*Enjoyment: the level of enjoyment involved in using the system", "Assistance: the level of perceived usefulness and helpfulness of the system in assisting users to conduct real tasks such as remote monitoring or distance learning".*

In order to compare other possible solutions with our real time best view acquisition system, three scenarios are tested. First is the *single camera area zoom scenario* which is a typical function provided by most of video camera vendors. To approximate best view acquisition using this scenario, the user has to first physically select the right camera and then apply a zoom-in function with their desired camera. Second is the *multiple cameras manual adjustment scenario* which is a function offered by our system (see section 4.1 The user interface). To obtain the best view of the object of interest, users can use the camera control panel in our system or screen-click to adjust every camera to the best position and then select the best camera view using the radio button (see Figure 4.1(b)). Third is the *automatic best view acquisition scenario* provided by our system. To obtain the best view, the user first select the object of interest by dragging an

rectangle around the object and then click the button “best view” (see Figure 4.1(b)). An example of results obtained from above three compared scenarios (conducted by one user) is shown in Figure 4.6.

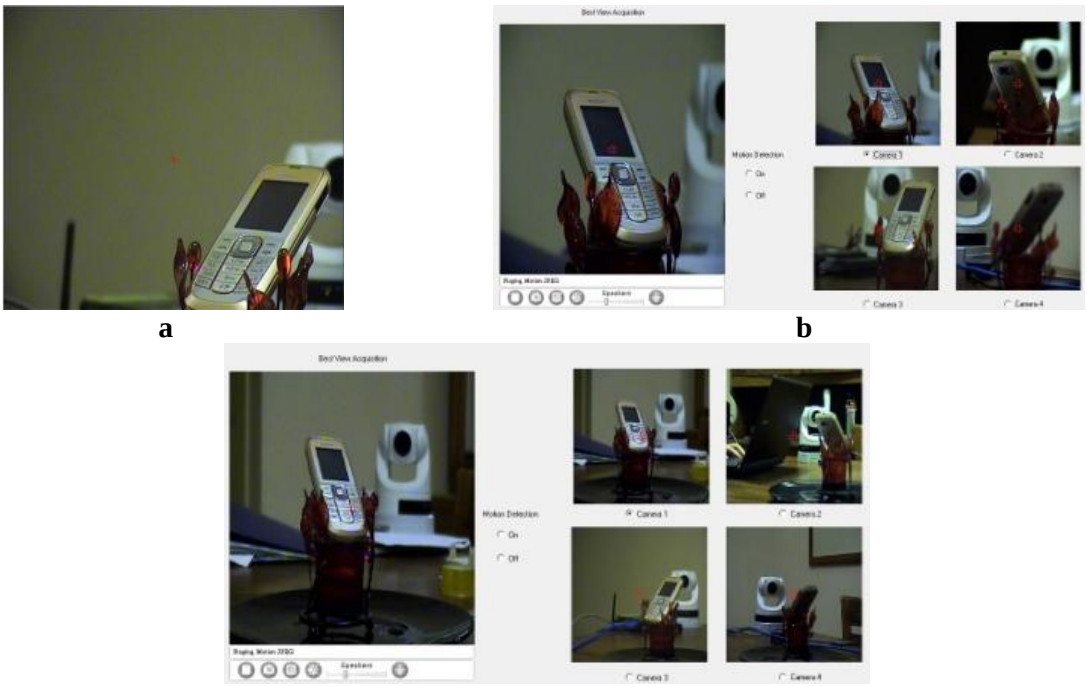


Figure 4.6 Best view acquisition results of three scenarios
a. Single camera area zoom-in; b. Multiple cameras manual adjustment;
c. Automatic best view acquisition.

10 users are invited to participate our test on above three scenarios and give their five-scale scores (i.e. 1: bad, 2: poor, 3: fair, 4: good, 5: excellent) on the five criterion stated above (ease of use, enjoyment, interactivity, assistance, and subjective-consistence). The test result is shown in Figure 4.7.

From Figure 4.7(b), we can see that our system demonstrates a improved quality of experience in “interactivity” and “subjective-consistence” provided by manual best view acquisition (scenario 2) and “ease of use”, “enjoyment” and “assistance” provided by automatic best view acquisition (scenario 3) compared with basic area zoom-in solution (scenario 1) offered by video camera’ venders. Also it can be seen from Figure 4.7(a), our automatic best view acquisition approach

(scenario 3) has around 10% "Poor" rating under the criteria interactivity and subjective-consistence, which probably because users may wish to have more opportunities to actually operate the cameras. This can be noticed from Figure 4.7(a) column 3 and 5: more than 40% "Excellent" ratings were given to the interactivity and subjective-consistence criteria for manual adjusting cameras (scenario 2) offered by our system.

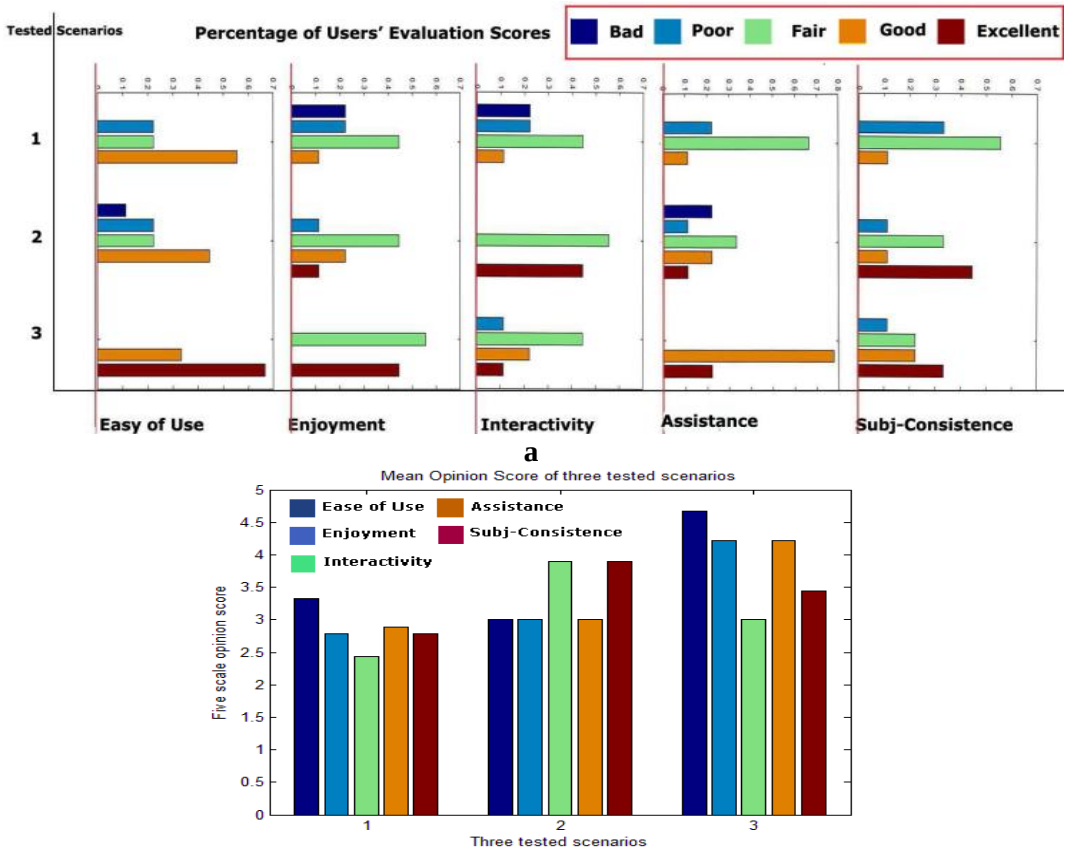


Figure 4.7 System QoE evaluation results
a. Percentage of users' evaluation scores for three scenarios;
b. Mean opinion score of three scenarios on five criteria;

In addition, scenario 3 is rated as the highest level of "ease of use" with more than 60% "Excellent" rating, "enjoyment" and "assistance" with around 80% "Good" rating (Figure 4.7(a) column 1, 2 and 4). Overall, we can conclude, compared with traditional cameras' vendors' offering, our system with combined features in tele-operation of multiple cameras and real time

best view acquisition is able to improve the quality of user experience in a few representative dimensions in QoS and QoE construct [73].

4.6. Discussion

In our experiment we found that Viewpoint Saliency score can be affected by strong texture information of the object. And we still need to improve our algorithm for acquiring the best view for humans. Furthermore, template matching results can affect our final results in the system. For complex objects that have largely different views from different angles, simple template matching may not be accuracy, instead pair-wised template matching for propagating the SIFT [37] key points between cameras' views should be applied. Finally, for moving objects, we still feel our current algorithm is not efficient and accurate enough to acquire objects with continuous or fast motion; new descriptor may be included in the Viewpoint Saliency definition. Also, when applying the Viewpoint Saliency metric, we have assumed that the scale does not change for all cameras' views which means that we do not fully utilize the zoom parameter of the cameras, this can also be seen from the comparison between manual (Figure 4.6(b)) and automatic acquisition (Figure 4.6 (c)) .

5. Conclusions

In this chapter, we conclude by first summarizing the overall work and major contributions of this thesis and then briefly outlining the possible future directions to improve and extend our current work.

5.1. Summary and contributions

Aiming to improve the QoE of real time applications for remote communication or education in cyber physical environments by providing users the best views of object(s) of interest, in this thesis, we first propose a new image-based viewpoint quality metric, Viewpoint Saliency (VS), for evaluating view qualities of captured cyber-physical environments. Based on VS, we propose a novel scheme to first map the 3D based viewpoint selection into 2D space and then control multiple cameras to obtain the best view upon the user's selection. Since the Viewpoint Saliency measure is purely image-based, 3D model reconstruction is not required. And then we map the real time best view selection and acquisition problem to a "Best Quality Least Effort" task on a graph formed by available views of an object and model it as a finite cameras state transition problem for energy minimization. Finally, the real time best view selection system is implemented on VC++ platform with multiple IP network cameras, and it demonstrates that our proposed approach is indeed feasible and effective for remotely acquiring the best views in cyber-physical environments via Internet. In addition, a user study of the system has demonstrated the improved QoE provided by the system.

The contributions of this thesis can be summarized as follows: first, an image based viewpoint evaluation metric, *Viewpoint Saliency,* is developed and tested and compared with previous 3D

based metrics; second, an energy minimization based camera control algorithm is proposed for acquiring the best view(s) of object(s) of interest to with the goal of "Best Quality Least Effort"; third, a system which supports remote best view selection and acquisition via Internet is implemented and tested with four IP network cameras on VC++ platform.

5.2. Future work

Although we have made some progress and provided an approach for the problem of "real time best view selection in cyber-physical environments", there are still many aspects that can be improved and extended based upon the current solution.

To be specific, in the future, we wish to improve the current Viewpoint Saliency metric to allow for better viewpoint quality evaluation of general objects where the zoom parameters of cameras can be fully utilized; meanwhile, we want to improve our current results especially for humans and moving objects. Furthermore, we wish to provide solutions to the situation that initially none of the cameras are at the positions where good views of the object of interest can be captured. In addition, in order to make the "best view "selection and acquisition feasible for World Wide Web based applications such as distance learning or remote monitoring, we wish to develop mechanism that allows multiple access to our current system and supports multiple users to obtain the best views of their own objects of interest.

The following paragraphs seek to identify above aspects and further analyze the underlying challenges for each of them in details.

(1) Fully utilize the zoom parameters of cameras to allow for better and more flexible viewpoint quality evaluation.

    The challenges of achieving this aspect including: (a) Change the definition of *VS* to incorporate viewpoint evaluation across different scales of object. (b) Need better segmentation and recognition of objects across different camera views. (c) Need camera calibration to estimate the size of objects of interest. (d) Need precise cameras control for optimal zoom.

(2) Refine *Viewpoint Saliency* (*VS*) measure to improve the viewpoint quality evaluation results for humans and moving objects

The challenges of achieving this aspect including: (a) New descriptors in *VS* definition need to be developed for evaluating viewpoint quality of human beings. (b) *VS* definition need to be refined to include "motion" feature for dealing with moving objects.

(3) Provide solutions to the situation that initially none of the cameras are at the positions where good views of the object of interest can be captured.

The challenges of achieving this aspect including: (a) Need better segmentation and recognition of objects across different camera views. (b) Need an algorithm to estimate the optimal initial camera positions for goods views of object to be captured. (c) Especially for cameras that are able to move, a control scheme is needed for moving the cameras to the optimal initial positions.

(4) Develop mechanism that allows multiple accesses to our current system and supports multiple users for best view selection via World Wide Web.

The challenges of achieving this aspect including: (a) Need an appropriate voting mechanism for handling multiple accesses to the system, i.e. decide when and who to serve, and who should wait. (b) Need an authentication mechanism to limit the administrating levels of users. (c) Need a protection/security mechanism for preventing adversarial users to maliciously manipulate cameras via WWW. (d) Need to investigate a distributed system architecture for handling the computational load in a scalable fashion.

Bibliography

[1] S. Ahmad, “VISIT: A neural model of covert attention,” Advances in Neural Information Processing Systems, Vol.4, p.420-427, San Mateo, CA: Morgan Kaufmann, 1991.

[2] P. Anjin, K. Jungwhan, M. Seungki, Y. Sungju, J. Keechul, "Graph Cuts-Based Automatic Color Image Segmentation," dicta, pp.564-571, 2008 Digital Image Computing: Techniques and Applications, 2008.

[3] P. Barral, G. Dorme, D. Plemenos “Visual understanding of a scene by automatic movement of a camera.” International Conference GraphiCon’99 (Aug-Sept 1999) Moscow Russia.

[4] A. Badano, MJ Flynn, J. Kanicki, “High fidelity medical imaging displays”, Bellingham, WA: SPIE Press, 2004.

[5] R. E. Blahut, “Principles and Practice of Information Theory,” Addison-Wesley, 1987.

[6] G. S. Bong, P. So-Y, J. L. Ju, “Fast and robust template matching algorithm in noisy image”, International Conference on Control, Automation and Systems, 2007. ICCAS’07, pp. 6-9, 17-20 Oct. 2007.

[7] N.D.B. Bruce, J. K. Tsotsos, “Saliency Based on Information Maximization”, Advances in Neural Information Processing Systems, 18, pp. 155-162, June 2006.

[8] N.D.B. Bruce, J. K. Tsotsos, “Saliency, Attention, and Visual Search: An Information Theoretic Approach”, Journal of Vision 9:3, p1-24, 2009.

[9] H. Barrett, J. Yao, J. Rolland, and K. Myers, "Model observers for assessment of image quality," proceedings of National Academy of Science of the USA, 90, pp. 9758-9756, Feb. 1993.

[10] C. I. Connolly, "The Determination of Next Best Views," IEEE International Conference on Robotics and Automation, pp. 432-435, Mar 1985.

[11] K. T. Chen, C. C. Wu, Y. C. Chang, C. L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. ACM International Conference on Multimedia, pp.491-500, Beijing, China, 2009.

[12] F. Deinzer, J. Denzler, H. Niemann, "Viewpoint selection-a classifier independent learning approach", proceedings of the 4th IEEE Southwest Symposium on Image Analysis and Interpretation, pp.209-213, 2-4 April. 2000.

[13] P. Datta, J. Li, J. Z. Wang, "Learning the consensus on visual quality for next-generation image management", proceedings of the 15th international conference on multimedia, pp 533-536, Augsburg, Germany, 2007.

[14] A. M. Eskicioglu, P. S. Fisher, "Image quality measures and their performance", IEEE transaction on communications, vol. 43, No.12, pp.2959 – 2965, Dec. 2005.

[15] P. Eli, "Contrast in complex images", J. Opt. Soc. Am. Vol. 7, Issue 10, pp. 2032-2040. 1990.

[16] M. Feixas, M. Sbert, F. Gonzalez "A unified information- theoretic framework for viewpoint selection and mesh saliency," ACM Transactions on Applied Perception, 2008.

[17] Goshtasby, Ardeshir, "Template matching in rotated images", IEEE Transactions on pattern analysis and machine intelligence, Volume PAMI-7, Issue 3, pp. 338-334, May 1985.

[18] K. Goldberg, S. Gentner, C. Sutter, J. Wiegley. The Mercury Project: A feasibility study for internet robots. the IEEE International Conference on Robotics and Automation, May 19-26, 1995, Nagoya, Japan.

[19] K. Han, Pencil sketch--Keeping memory of Beijing's Hutong: http://www.chinatoday.com/art/pencil.sketching.hutong/pencil_sketch_hutong_18.htm

[20] L. Itti, C. Koch, "Computational Modeling of Visual Attention", Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, Mar 2001.

[21] L. Itti, C. Koch, "Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems", Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA, Vol. 3644, pp. 473-82, Bellingham, WA:SPIE Press, Jan 1999.

[22] L. Itti, C. Koch, E. Niebur, "A model of Saliency- based Visual Attention for Rapid Scene Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp. 1254-1259, Nov 1998.

[23] W. James, "The principles of psychology," Harvard University Press, 1890.

[24] H. Y. Kim, S. A. Araújo, "Grayscale Template-Matching Invariant to Rotation, Scale, Translation, Brightness and Contrast," IEEE Pacific-Rim Symposium on Image and Video Technology, Lecture Notes in Computer Science, vol. 4872, pp. 100-113, 2007.

[25] F. Karel, "Eyetracking based approach to objective image quality assessment", Security Technology, 2008. ICCST 2008. 42nd Annual IEEE International Carnahan conference on, pp.371-376, 13-16 Oct. 2008.

[26] A. A. Khwaja, R. Goecke, "Image reconstruction from contrast information", Digital Image Computing: Techniques and Applications, 2008. DICTA'08 Digital Image, pp 226-233. 1-3 Dec. 2008.

[27] S. Kiss, A. Nijholt, " Viewpoint adaptation during navigation based on stimuli from the virtual environment", proceedings of the 8th international conference on 3D Web technology, Session 1, p. 19-26, Saint Malo, France, 2003.

[28] D. Lamming, Contrast Sensitivity. Chapter 5. In: Cronly-Dillon, J., Vision and Visual Dysfunction, Vol 5, London: Macmillan Press. 1991

[29] C. Li-Wei, C. Cheng-Chieh, and H. Yi-Ping "Content-Based Object Movie Retrieval by Use of Relevance Feedback", 4th international conference on image and video retrieval (CIVR) pp. 425-434, 2005.

[30] G. E. Legge, J. M. Foley, "Contrast masking in human vision", J. Opt. Soc. Am., Vol. 70, No.12, December 1980.

[31] J. Lin, "Divergence measures based on the Shannon entropy", IEEE Transactions on Information Theory, 37(1): 145-151, January 1991.

[32] S. Lee , G. J. Kim, S. Choi, Real-time tracking of visually attended objects in interactive virtual environments, Proceedings of the 2007 ACM symposium on Virtual reality software and technology, November 05-07, Newport Beach, California 2007.

[33] Q. Liu, D. Kimber, J. Foote, C. Liao. "Multichannel video/audio acquisition for immersive conferencing", Proc. IEEE Int. Conf. Multimedia Expo(ICME), July 2003.

[34] Q. Liu, D. Kimber, L. Wilcox, M. Cooper, J. Foote, J. Boreczky. "Managing a camera system to serve different video requests", Proc. IEEE Int. Conf. Multimedia Expo (ICME), vol.2, pp. 13-16, Lausanne, Switzerland, Aug. 2002.

[35] K. L. Low, A. Lastra, "An adaptive hierarchical next best view algorithm for 3D reconstruction of indoor scenes", 14th Pacific Conference on Computer Graphics and Applications (Pacific Graphics 2006) , Taipei, Taiwan, Oct. 2006.

[36] Y. Li and K. L. Low, “Automatic Registration of Color Images to 3D geometry”, 27th Computer Graphics International Conference (CGI 2009), Victoria, British Columbia, Canada, May 2009.

[37] D. G. Lowe, "Object recognition from local scale-invariant features", International Conference on Computer Vision, Corfu, Greece (September 1999), pp. 1150-1157.

[38] L. Li , M. Tao , H. Xian-Sheng , L. Shipeng, “ImageSense”, Proceeding of the 16th ACM international conference on Multimedia, Vancouver, British Columbia, Canada, October 26-31, 2008.

[39] C. J. B. Lambrecht, O. Verscheure, “Perceptual Quality Measure using a spatio-temporal model of the human visual system”, proceedings of the SPIE, vol. 2668, pp. 450-461, IEEE, 1996.

[40] C. H. Lee, , A. Varshney, D. W. Jacobs, “Mesh saliency”, ACM Transactions on Graphics (Proceedings 5 of SIGGRAPH’05) Vol. 24, No.3, 659-666.

[41] C. Maarten, P. V. Arjen, J. T. R. Marcel, “Exploiting positive and negative graded relevance assessments for content recommendation”, Algorithms and Models for the Web-Graph, Springer Berlin, pp155-166, 2009.

[42] J. L. Mannos, D. J. Sakrison, ``The Effects of a Visual Fidelity Criterion on the Encoding of Images", IEEE Transactions on Information Theory, pp. 525-535, Vol. 20, No 4, 1974.

[43] A. Mcnamara, “Exploring visual and automatic measures of perceptual fidelity in real and simulated imagery”, ACM Transaction on Applied Perception, Vol. 3, No. 3, pp217-238, July, 2006.

[44] N. A. Massios, R. B. Fisher, “A best next view selection algorithm incorporating a quality criterion,” Proceedings.of the Britsh Machine Vision Conference, 1998.

[45] A. Murata, H. Iwase, "Visual attention models-object-based theory of visual attention", proceedings of the IEEE international conference on system, man, and cybernetics, 1999. SMC'99, vol. 2, pp 60-65, Tokyo, Japan, 1999.

[46] R. S. Mosher. Industrial Manipulators. Scientific American, 211(4), 1964.

[47] S. Mata, L. Pastor, J. J. Aliaga, A. Rodriguez, "Incorporating visual attention into mesh simplification techniques", proceedings of the 4th symposium on applied perception in graphics and visualization , vol.253, pp. 134-134, Tubingen, Germany 2007.

[48] P. M. Moreira, L. P. Reis, A. A. Sousa, "Best multiple view selection for the visualization of urban rescue simulations", International Symposium CompIMAGE-Coimbra, Portugal, 20-21 October 2006.

[49] Y. F. Ma, H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," Proceedings of the 11th ACM international conference on Multimedia, pp.374-381, Berkeley, CA, USA, 2003.

[50] E. Niebur, C. Koch, "Computational architectures for attention," R. Parasuraman, (Ed.), The attentive brain, Cambridge, MA: MIT Press, pp. 163-186, 1998.

[51] W. Osberger, N. Bergmann, A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors", Proc. Image Processing, vol.3, pp. 414-418, 4-7 Oct. 1998.

[52] S. Omachi, M. Omachi, "Fast template matching with polynomials", IEEE Transactions on Image processing, vol. 16, Issue 8, pp. 2139-2149, Aug. 2007.

[53] C. Oprea, I. Pirnog, C. Paleologu, M. Udrea, "Perceptual video quality assessment based on salient region detection", 2009 fifth advanced international conference on telecommunications, AICT'09, p.232-236, 24-28 May 2009.

[54] J. Park, P. C. Bhat, A. C. Kak, “A Look-up table based approach for solving the camera selection problem in large camera networks”, workshop on distributed smart cameras in conjunction with ACM SenSys’ 06, 2006.

[55] N. Querhani, H. Hugli, “Computing visual attention from scene depth”, proceedings of the international conference on pattern recognition, vol. 1, pp.375-378, Washington DC, USA, 2000.

[56] J. Radun, T. Leisti, J. Hakkinen, H. Ojanen, J. Olives, T. Vuori, G. Nyman, “Content and quality: Interpretation-based Estimation of image quality”, ACM Transactions on Applied Perception, Vol. 4, No.4, Article 21, Jan. 2008.

[57] D. Rouse, S. S. Hemami, "Understanding and Simplifying the Structural Similarity Metric," presented at IEEE Intl. Conf. of Image Proc. (ICIP) San Diego, CA, October 2008.

[58] D. Rouse, R. Pepion, S. S. Hemami, P. L. Callet, "Image Utility Assessment and a Relationship with Image Quality Assessment," Proc. SPIE Vol. 7240, Human Vision and Electronic Imaging, San Jose, CA, January 2009.

[59] H. R. Sheikh, A. C. Bovik, “Image information and visual quality,” IEEE Trans. Image Processing 15, pp. 430-444, Feb. 2006.

[60] D. Song, K. Goldberg, “Approximate Algorithms for a Collaboratively Controlled Robotic Camera”, IEEE Transactions on Robotics. Vol. 23, No. 5, Oct 2007.

[61] D. Song, N. Qin, K. Goldberg, “Systems, Control Models, and Codec for Collaborative Observation of Remote Environments with an Autonomous Networked Robotic Camera”, Autonomous Robots. Vol. 24, No. 4. May 2008.

[62] L. Snidaro, R. Niu, P. K. Varshney, G. L. Foresti, "Automatic camera selection and fusion for outdoor surveillance under changing weather conditions", proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03), 2003.

[63] D. Sokolov, D. Plemenos, "Viewpoint quality and scene understanding", In Mudge, M., Ryan, N., and Scopigno, R., editors, VAST 2005: Eurographics Symposium Proceedings., pp 67-73, ISTI-CNR Pisa, Italy. Eurographics Association.

[64] M. Sbert, D. Plemenos, M. Feixas, F. Gonzalez, "Viewpoint quality: measures and applications", Computational Aesthetics in Graphics, Visualization and Imaging, 185-192. 2005.

[65] N. Vaswani, R. Chellappa "Best view selection and compression of moving objects in IR sequences," Proceedings of the Acoustics, Speech, and Signal Processing, 2001. on IEEE International Conference, Vol.03, pp.1617-1620, 2001.

[66] P. P. Vazquez, M. Feixas, M. Sbert, W. Heidrich, "Viewpoint selection using viewpoint entropy," Proceedings of Vision, Modeling and Visualization 2001(Stuttgart, Germany, November 2001),Ertl T., GirodB., Greiner G., Niemann H., Seidel H.-P., (Eds.), pp.273-280. Stuttgart, Germany.

[67] P. Vazquez, M. Feixas, M. Sbert, W. Heidrich, "Automatic View Selection Using Viewpoint Entropy and its Application to Image-Based Rendering", Computer Graphics Forum, 22(4), pp. 689-700, 2003.

[68] P. P. Vazquez, M. Feixas, M. Sbert, A. Llobet, "Viewpoint entropy: A New Tool for Obtaining Good Views for Molecules," Data Visualization 2002 (Eurographics /IEEE TCVG Symposium Proceedings). Barcelona, Spain May 27-29, 2002.

[69] P. P. Vázquez, Sbert M.,“Automatic Indoor Scene Exploration” Proc. of 6th International Conference on Computer Graphics and Artificial Intelligence, pp. 13-24, 2003.

[70] P. Vazquez, M. Sbert, “Fast adaptive selection of best views”, International Conference on Computational Science and its Applications, ICCSA'2003. (LNCS 2669), 2003.

[71] P. P. Vázquez, M. Sbert, “On the fly detection of best views using graphics hardware,” 4th IASTED International Conference on Visualization, Image, and Image Processing, VIIP 2004.

[72] I. Viola, M. Feixas, M. Sbert, M. E. Groller, “Importance –Driven Focus of Attention”, IEEE Transactions on Visualization and Computer Graphics”, vol. 12, Issue 5, pp.933-940, Sept.-Oct. 2006.

[73] W. Wu, A. Arefin, R. Rivas., K. Nahrstedt, R. M. Sheppard, Z. Yang. Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. ACM International Conference on Multimedia, pp.481-490, Beijing, China, 2009.

[74] Z. Wang, A. C. Bovik, L. Lu, “Why is image quality assessment so difficulty?”, IEEE international conference on acoustics, speech, and signal processing, may 2002.

[75] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, “Image quality assessment: From error measurement to structural similarity”, IEEE Transaction on Image Processing, 13(4), pp. 600-612. April 2004.

[76] X. Wei, J. Li, G. Chen, “An image quality estimation model based on HSV”, TENCON 2006. 2006 IEEE Region 10 Conference, pp.1-4, 14-17 Nov. 2006.

[77] C. Y. Wu, J. J. Leou, H. Y. Chen, “Visual attention region determination using low-level features”, IEEE international Symposium on Circuits and Systems, ISCAS, pp. 3178-3181, May 24 2009.

[78] Z. Xenophon, D. Kostas, "Multi-Camera Reconstruction based on Surface Normal Estimation and Best Viewpoint Selection," 3dpvt, pp.733-740, Second International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT'04), 2004.

[79] J. You, A. Perkis, M. M. Hannuksela, M. Gabbouj. Perceptual quality assessment based on visual attention analysis. ACM International Conference on Multimedia, pp.561-564, Beijing, China, 2009.

[80] The Free Dictionary by farlex (medical dictionary), pulvinar:

http://medical-dictionary.thefreedictionary.com/pulvinar